

**APPLIED
MATHEMATICS
SECTION
STATISTICS**

Student Files

Guidance Notes

1. The work that you will produce during your Course will build up into an important document known as the Student File.

2. It is essential for your success on the course, that your file is well organised, kept up to date, and is well presented.

3. Your file will be *looked at* by a number of different people

- you during revision and assignment writing
- your tutors
- moderators when assessing portfolios
- inspectors

4. The nature of your file will depend to a large extent on the type of course you are following, e.g. GNVQ and BTEC courses will have a large number of assignments within the file, whilst A Level courses will have more factual notes many of which have been made by the student. The following notes, however, are designed to help you organise your file more effectively, no matter what course you are on.

5. Here are some Do's when building up your course file :

- Make a "contents" list at the beginning of your file, and keep all work in its correct section.
- When given handouts or references try to "personalise" them by adding your own notes or highlight key points.
- Try to remember to add a date and to initial any material given to you by your tutor.
- Skim-read reference material or handouts and select key words or phrases that summarise the main points of the text. Develop your own technique for doing this e.g. highlighter pen, underline in red etc.
- Make sure that your file contains work that you have produced. Summaries of reference texts or articles that you have been directed to should be placed in your file, as well as your record of assignments, homeworks and other forms of assessment.
- Keep all records of assessment carefully in your file and note your tutors comments.
- Look after your file.

Sampling (statistics)

From Wikipedia, the free encyclopedia

Sampling is that part of statistical practice concerned with the selection of an unbiased or random subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern, especially for the purposes of making predictions based on statistical inference. Sampling is an important aspect of data collection.

Researchers rarely survey the entire population for two reasons (Adér, Mellenbergh, & Hand, 2008): the cost is too high, and the population is dynamic in that the individuals making up the population may change over time. The three main advantages of sampling are that the cost is lower, data collection is faster, and since the data set is smaller it is possible to ensure homogeneity and to improve the accuracy and quality of the data.

Each observation measures one or more properties (such as weight, location, color) of observable bodies distinguished as independent objects or individuals. In survey sampling, survey weights can be applied to the data to adjust for the sample design. Results from probability theory and statistical theory are employed to guide practice. In business and medical research, sampling is widely used for gathering information about a population.^[1]

The Sampling Process

The sampling process comprises several stages:

- Defining the population of concern
- Specifying a sampling frame, a set of items or events possible to measure
- Specifying a sampling method for selecting items or events from the frame
- Determining the sample size
- Implementing the sampling plan
- Sampling and data collecting
- Reviewing the sampling process

Population definition

Successful statistical practice is based on focused problem definition. In sampling, this includes defining the population from which our sample is drawn. A population can be defined as including all people or items with the characteristic one wishes to understand. Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.

Sometimes that which defines a population is obvious. For example, a manufacturer needs to decide whether a batch of material from production is of high enough quality

to be released to the customer, or should be sentenced for scrap or rework due to poor quality. In this case, the batch is the population.

Although the population of interest often consists of physical objects, sometimes we need to sample over time, space, or some combination of these dimensions. For instance, an investigation of supermarket staffing could examine checkout line length at various times, or a study on endangered penguins might aim to understand their usage of various hunting grounds over time. For the time dimension, the focus may be on periods or discrete occasions.

In other cases, our 'population' may be even less tangible. For example, Joseph Jagger studied the behaviour of roulette wheels at a casino in Monte Carlo, and used this to identify a biased wheel. In this case, the 'population' Jagger wanted to investigate was the overall behaviour of the wheel (i.e. the probability distribution of its results over infinitely many trials), while his 'sample' was formed from observed results from that wheel. Similar considerations arise when taking repeated measurements of some physical characteristic such as the electrical conductivity of copper.

This situation often arises when we seek knowledge about the cause system of which the *observed* population is an outcome. In such cases, sampling theory may treat the observed population as a sample from a larger 'superpopulation'. For example, a researcher might study the success rate of a new 'quit smoking' program on a test group of 100 patients, in order to predict the effects of the program if it were made available nationwide. Here the superpopulation is "everybody in the country, given access to this treatment" - a group which does not yet exist, since the program isn't yet available to all.

Note also that the population from which the sample is drawn may not be the same as the population about which we actually want information. Often there is large but not complete overlap between these two groups due to frame issues etc. (see below). Sometimes they may be entirely separate - for instance, we might study rats in order to get a better understanding of human health, or we might study records from people born in 2008 in order to make predictions about people born in 2009.

Time spent in making the sampled population and population of concern precise is often well spent, because it raises many issues, ambiguities and questions that would otherwise have been overlooked at this stage.

Sampling frame

In the most straightforward case, such as the sentencing of a batch of material from production (acceptance sampling by lots), it is possible to identify and measure every single item in the population and to include any one of them in our sample. However, in the more general case this is not possible. There is no way to identify all rats in the set of all rats. Where voting is not compulsory, there is no way to identify which people will actually vote at a forthcoming election (in advance of the election).

These imprecise populations are not amenable to sampling in any of the ways below and to which we could apply statistical theory.

As a remedy, we seek a *sampling frame* which has the property that we can identify every single element and include any in our sample.^{III} The most straightforward type of frame is a list of elements of the population (preferably the entire population) with appropriate contact information. For example, in an opinion poll, possible sampling frames include:

- Electoral register
- Telephone directory

Not all frames explicitly list population elements. For example, a street map can be used as a frame for a door-to-door survey; although it doesn't show individual houses, we can select streets from the map and then visit all houses on those streets. (One advantage of such a frame is that it would include people who have recently moved and are not yet on the list frames discussed above.)

The sampling frame must be representative of the population and this is a question outside the scope of statistical theory demanding the judgment of experts in the particular subject matter being studied. All the above frames omit some people who will vote at the next election and contain some people who will not; some frames will contain multiple records for the same person. People not in the frame have no prospect of being sampled. Statistical theory tells us about the uncertainties in extrapolating from a sample to the frame. In extrapolating from frame to population, its role is motivational and suggestive.

To the scientist, however, representative sampling is the only justified procedure for choosing individual objects for use as the basis of generalization, and is therefore usually the only acceptable basis for ascertaining truth.

—Andrew A. Marino¹²

It is important to understand this difference to steer clear of confusing prescriptions found in many web pages.

In defining the frame, practical, economic, ethical, and technical issues need to be addressed. The need to obtain timely results may prevent extending the frame far into the future.

The difficulties can be extreme when the population and frame are disjoint. This is a particular problem in forecasting where inferences about the future are made from historical data. In fact, in 1703, when Jacob Bernoulli proposed to Gottfried Leibniz the possibility of using historical mortality data to predict the probability of early death of a living man, Gottfried Leibniz recognized the problem in replying:

Nature has established patterns originating in the return of events but only for the most part. New illnesses flood the human race, so that no matter how many experiments you have done on corpses, you have not thereby imposed a limit on the nature of events so that in the future they could not vary.

—Gottfried Leibniz

Kish posited four basic problems of sampling frames:

1. Missing elements: Some members of the population are not included in the frame.
2. Foreign elements: The non-members of the population are included in the frame.
3. Duplicate entries: A member of the population is surveyed more than once.
4. Groups or clusters: The frame lists clusters instead of individuals.

A frame may also provide additional 'auxiliary information' about its elements; when this information is related to variables or groups of interest, it may be used to improve survey design. For instance, an electoral register might include name and sex; this information can be used to ensure that a sample taken from that frame covers all demographic categories of interest. (Sometimes the auxiliary information is less explicit; for instance, a telephone number may provide some information about location.)

Having established the frame, there are a number of ways for organizing it to improve efficiency and effectiveness.

It's at this stage that the researcher should decide whether the sample is in fact to be the whole population and would therefore be a census.

Probability and nonprobability sampling

A probability sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Example: We want to estimate the total income of adults living in a given street. We visit each household in that street, identify all adults living there, and randomly select one adult from each household. (For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income. People living on their own are certain to be selected, so we simply add their income to our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (In effect, the person who is selected from that household is taken as representing the person who isn't selected.)

In the above example, not everybody has the same probability of selection; what makes it a probability sample is the fact that each person's probability is known. When every element in the population *does* have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

Probability sampling includes: Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or

Multistage Sampling. These various ways of probability sampling have two things in common:

1. Every element has a known nonzero probability of being sampled and
2. involves random selection at some point.

Nonprobability sampling is any sampling method where some elements of the population have *no* chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, nonprobability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Example: We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a nonprobability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities.

Nonprobability Sampling includes: Accidental Sampling, Quota Sampling and Purposive Sampling. In addition, nonresponse effects may turn *any* probability design into a nonprobability design if the characteristics of nonresponse are not well understood, since nonresponse effectively modifies each element's probability of being sampled.

Sampling methods

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

- Nature and quality of the frame
- Availability of auxiliary information about units on the frame
- Accuracy requirements, and the need to measure accuracy
- Whether detailed analysis of the sample is expected
- Cost/operational concerns

Simple random sampling

In a simple random sample ('SRS') of a given size, all such subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection: the frame is not subdivided or partitioned. Furthermore, any given *pair* of elements has the same chance of selection as any other such pair (and similarly for triples, and so on). This minimises bias and simplifies analysis of results. In

particular, the variance between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

However, SRS can be vulnerable to sampling error because the randomness of the selection may result in a sample that doesn't reflect the makeup of the population. For instance, a simple random sample of ten people from a given country will on average produce five men and five women, but any given trial is likely to overrepresent one sex and underrepresent the other. Systematic and stratified techniques, discussed below, attempt to overcome this problem by using information about the population to choose a more representative sample.

SRS may also be cumbersome and tedious when sampling from an unusually large target population. In some cases, investigators are interested in research questions specific to subgroups of the population. For example, researchers might be interested in examining whether cognitive ability as a predictor of job performance is equally applicable across racial groups. SRS cannot accommodate the needs of researchers in this situation because it does not provide subsamples of the population. Stratified sampling, which is discussed below, addresses this weakness of SRS.

Simple random sampling is always an FPS design, but not all FPS designs are simple random sampling.

Systematic sampling

Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every k th element from then onwards. In this case, k =(population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the k th element in the list. A simple example would be to select every 10th name from the telephone directory (an every 10th sample, also referred to as 'sampling with a skip of 10').

As long as the starting point is randomized, systematic sampling is a type of probability sampling. It is easy to implement and the stratification induced can make it efficient, if the variable by which the list is ordered is correlated with the variable of interest. Every 10th sample is especially useful for efficient sampling from databases.

Example: Suppose we wish to sample people from a long street that starts in a poor district (house #1) and ends in an expensive district (house #1000). A simple random selection of addresses from this street could easily end up with 100 many from the high end and too few from the low end (or vice versa), leading to an unrepresentative sample. Selecting (e.g.) every 10th street number along the street ensures that the sample is spread evenly along the length of the street, representing all of these districts. (Note that if we always start at house #1 and end at #991, the sample is slightly biased towards the low end, by randomly selecting the start between #1 and #10, this bias is eliminated.)

However, systematic sampling is especially vulnerable to periodicities in the list. If periodicity is present and the period is a multiple or factor of the interval used, the sample is especially likely to be unrepresentative of the overall population, making the scheme less accurate than simple random sampling.

Example: Consider a street where the odd-numbered houses are all on the north (expensive) side of the road, and the even-numbered houses are all on the south (cheap) side. Under the sampling scheme given above, it is impossible to get a representative sample: either the houses sampled will all be from the odd-numbered, expensive side, or they will all be from the even-numbered, cheap side.

Another drawback of systematic sampling is that even in scenarios where it is more accurate than SRS, its theoretical properties make it difficult to quantify that accuracy. (In the two examples of systematic sampling that are given above, much of the potential sampling error is due to variation between neighbouring houses - but because this method never selects two neighbouring houses, the sample will not give us any information on that variation.)

As described above, systematic sampling is an EPS method, because all elements have the same probability of selection (in the example given, one in ten). It is not 'simple random sampling' because different subsets of the same size have different selection probabilities - e.g. the set $\{4, 14, 24, \dots, 94\}$ has a one-in-ten probability of selection, but the set $\{4, 13, 24, 34, \dots\}$ has zero probability of selection.

Systematic sampling can also be adapted to a non-EPS approach; for an example, see discussion of PPS samples below.

Stratified sampling

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected^[3]. There are several potential benefits to stratified sampling.

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.

Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group's size in the population.

Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).

Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates. Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling.

A stratified sampling approach is most effective when three conditions are met

1. Variability within strata are minimized
2. Variability between strata are maximized
3. The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

Advantages over other sampling methods

1. Focuses on important subpopulations and ignores irrelevant ones.
2. Allows use of different sampling techniques for different subpopulations.
3. Improves the accuracy/efficiency of estimation.
4. Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

Disadvantages

1. Requires selection of relevant stratification variables which can be difficult.
2. Is not useful when there are no homogeneous subgroups.
3. Can be expensive to implement.

Poststratification

Stratification is sometimes introduced after the sampling phase in a process called "poststratification".^[3] This approach is typically implemented due to a lack of prior knowledge of an appropriate stratifying variable or when the experimenter lacks the necessary information to create a stratifying variable during the sampling phase. Although the method is susceptible to the pitfalls of post hoc approaches, it can provide several benefits in the right situation. Implementation usually follows a simple random sample. In addition to allowing for stratification on an ancillary variable, poststratification can be used to implement weighting, which can improve the precision of a sample's estimates.^[3]

Choice-based sampling is one of the stratified sampling strategies. In choice-based sampling^[4], the data are stratified on the target and a sample is taken from each strata so that the rare target class will be more represented in the sample. The model is then built on this biased sample. The effects of the input variables on the target are often estimated with more precision with the choice-based sample even when a smaller overall sample size is taken, compared to a random sample. The results usually must be adjusted to correct for the oversampling.

Probability proportional to size sampling

In some cases the sample designer has access to an "auxiliary variable" or "size measure", believed to be correlated to the variable of interest, for each element in the population. This data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above.

Another option is probability-proportional-to-size (PPS) sampling, in which the selection probability for each element is set to be proportional to its size measure, up to a maximum of 1. In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawbacks of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections. To address this problem, PPS may be combined with a systematic approach.

Example: Suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490 students respectively (total 1500 students), and we want to use student population as the basis for a PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 ($= 150 + 180$), the third school 331 to 530, and so on to the last school (1011 to 1500). We then generate a random start between 1 and 500 (equal to $1500/3$) and count through the school populations by multiples of 500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137, i.e. the first, fourth, and sixth schools.

The PPS approach can improve accuracy for a given sample size by concentrating sample on large elements that have the greatest impact on population estimates. PPS sampling is commonly used for surveys of businesses, where element size varies greatly and auxiliary information is often available - for instance, a survey attempting to measure the number of guest-nights spent in hotels might use each hotel's number of rooms as an auxiliary variable. In some cases, an older measurement of the variable of interest can be used as an auxiliary variable when attempting to produce more current estimates.

Cluster sampling

Sometimes it is cheaper to 'cluster' the sample in some way e.g. by selecting respondents from certain areas only, or certain time-periods only. (Nearly all samples are in some sense 'clustered' in time - although this is rarely taken into account in the analysis.)

Cluster sampling is an example of 'two-stage sampling' or 'multistage sampling': in the first stage a sample of areas is chosen; in the second stage a sample of respondents within those areas is selected.

This can reduce travel and other administrative costs. It also means that one does not need a sampling frame listing all elements in the target population. Instead, clusters can be chosen from a cluster-level frame, with an element-level frame created only for the selected clusters. Cluster sampling generally increases the variability of sample estimates above that of simple random sampling, depending on how the clusters differ between themselves, as compared with the within-cluster variation.

Nevertheless, some of the disadvantages of cluster sampling are the reliance of sample estimate precision on the actual clusters chosen. If clusters chosen are biased in a certain way, inferences drawn about population parameters from these sample estimates will be far off from being accurate.

Multistage sampling Multistage sampling is a complex form of cluster sampling in which two or more levels of units are embedded one in the other. The first stage consists of constructing the clusters that will be used to sample from. In the second stage, a sample of primary units is randomly selected from each cluster (rather than using all units contained in all selected clusters). In following stages, in each of those selected clusters, additional samples of units are selected, and so on. All ultimate units (individuals, for instance) selected at the last step of this procedure are then surveyed.

This technique, thus, is essentially the process of taking random samples of preceding random samples. It is not as effective as true random sampling, but it probably solves more of the problems inherent to random sampling. Moreover, it is an effective strategy because it banks on multiple randomizations. As such, it is extremely useful.

Multistage sampling is used frequently when a complete list of all members of the sample units in all selected clusters, multistage sampling avoids the large, and perhaps unnecessary, costs associated traditional cluster sampling.

Matched random sampling

A method of assigning participants to groups in which pairs of participants are first matched on some characteristic and then individually assigned randomly to groups.^[5]

The procedure for matched random sampling can be briefed with the following contexts,

1. Two samples in which the members are clearly paired, or are matched explicitly by the researcher. For example, IQ measurements or pairs of identical twins.
2. Those samples in which the same attribute, or variable, is measured twice on each subject, under different circumstances. Commonly called repeated measures. Examples include the times of a group of athletes for 1500m before and after a week of special training; the milk yields of cows before and after being fed a particular diet.

Quota sampling

In quota sampling, the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.

It is this second step which makes the technique one of non-probability sampling. In quota sampling the selection of the sample is non-random. For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for many years

Convenience sampling

Convenience sampling (sometimes known as grab or opportunity sampling) is a type of nonprobability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, a sample population selected because it is readily available and convenient. The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough. For example, if the interviewer was to conduct such a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey was to be conducted at different times of day and several times per week. This type of sampling is most useful for pilot testing. Several important considerations for researchers using convenience samples include:

1. Are there controls within the research design or experiment which can serve to lessen the impact of a non-random, convenience sample whereby ensuring the results will be more representative of the population?
2. Is there good reason to believe that a particular convenience sample would or should respond or behave differently than a random sample from the same population?
3. Is the question being asked by the research one that can adequately be answered using a convenience sample?

In social science research, snowball sampling is a similar technique, where existing study subjects are used Line- to recruit more subjects into the sample.

Line intercept sampling

Line-intercept sampling is a method of sampling elements in a region whereby an element is sampled if a chosen line segment, called a "transect", intersects the element.

Panel sampling

Panel sampling is the method of first selecting a group of participants through a random sampling method and then asking that group for the same information again several times over a period of time. Therefore, each participant is given the same survey or interview at two or more time points; each period of data collection is called a "wave". This sampling methodology is often chosen for large scale or nation-wide studies in order to gauge changes in the population with regard to any number of variables from chronic illness to job stress to weekly food expenditures. Panel sampling can also be used to inform researchers about within-person health changes due to age or help explain changes in continuous dependent variables such as spousal interaction. There have been several proposed methods of analyzing panel sample data, including MANOVA, growth curves, and structural equation modeling with lagged effects. For a more thorough look at analytical techniques for panel data, see Johnson (1995).

Event sampling methodology

Event sampling methodology (ESM) is a new form of sampling method that allows researchers to study ongoing experiences and events that vary across and within days in its naturally-occurring environment. Because of the frequent sampling of events inherent in ESM, it enables researchers to measure the topology of activity and detect the temporal and dynamic fluctuations of work experiences. Popularity of ESM as a new form of research design increased over the recent years because it addresses the shortcomings of cross-sectional research, where once unable to, researchers can now detect intra-individual variances across time. In ESM, participants are asked to record their experiences and perceptions in a paper or electronic diary.

There are three types of ESM:

1. Signal contingent – random beeping notifies participants to record data. The advantage of this type of ESM is minimization of recall bias.
2. Event contingent – records data when certain events occur
3. Interval contingent – records data according to the passing of a certain period of time

ESM has several disadvantages. One of the disadvantages of ESM is it can sometimes be perceived as invasive and intrusive by participants. ESM also leads to possible self-selection bias. It may be that only certain types of individuals are willing to participate in this type of study creating a non-random sample. Another concern is related to participant cooperation. Participants may not be actually fill out their diaries at the specified times. Furthermore, ESM may substantively change the phenomenon being studied. Reactivity or priming effects may occur, such that repeated measurement may cause changes in the participants' experiences. This method of sampling data is also highly vulnerable to common method variance.^[6]

Further, it is important to think about whether or not an appropriate dependent variable is being used in an ESM design. For example, it might be logical to use ESM in order to answer research questions which involve dependent variables with a great deal of variation throughout the day. Thus, variables such as change in mood, change

in stress level, or the immediate impact of particular events may be best studied using ESM methodology. However, it is not likely that utilizing ESM will yield meaningful predictions when measuring someone performing a repetitive task throughout the day or when dependent variables are long-term in nature (coronary heart problems).

Replacement of selected units

Sampling schemes may be *without replacement* ('WOR' - no element can be selected more than once in the same sample) or *with replacement* ('WR' - an element may appear multiple times in the one sample). For example, if we catch fish, measure them, and immediately return them to the water before continuing with the sample, this is a WR design, because we might end up catching and measuring the same fish more than once. However, if we do not return the fish to the water (e.g. if we eat the fish), this becomes a WOR design.

Sample size

Formulas, tables, and power function charts are well known approaches to determine sample size.

Formulas

Where the frame and population are identical, statistical theory yields exact recommendations on sample size.^[7] However, where it is not straightforward to define a frame representative of the population, it is more important to understand the cause of variation are embraced in the frame. Large number of observations are of no value if major sources of variation are neglected in the study. In other words, it is taking a sample group that matches the survey category and is easy to survey. Bartlett, Kotlik, and Higgins (2001) published a paper titled *Organizational Research: Determining Appropriate Sample Size in Survey Research* Information Technology, Learning, and Performance Journal^[8] that provides an explanation of Cochran's (1977) Formulas. A discussion and illustration of sample size formulas, including the formula for adjusting the sample size for smaller populations, is included. A table is provided that can be used to select the sample size for a research problem based on three alpha levels and a set error rate.

Steps for using sample size tables

1. Postulate the effect size of interest, α , and β .
2. Check sample size table^[9]
1. Select the table corresponding to the selected α
2. Locate the row corresponding to the desired power
3. Locate the column corresponding to the estimated effect size.
4. The intersection of the column and row is the minimum sample size required.
- 5.

Sampling and data collection

Good data collection involves:

- Following the defined sampling process
- Keeping the data in time order
- Noting comments and other contextual events
- Recording non-responses

Most sampling books and papers written by non-statisticians focus only in the data collection aspect, which is just a small though important part of the sampling process.

Errors in sample surveys

Survey results are typically subject to some error. Total errors can be classified into sampling errors and non-sampling errors. The term "error" here includes systematic biases as well as random errors.

Sampling errors and biases

Sampling errors and biases are induced by the sample design. They include:

1. **Selection bias:** When the true selection probabilities differ from those assumed in calculating the results.
2. **Random sampling error:** Random variation in the results due to the elements in the sample being selected at random.

Non-sampling error

Non-sampling errors are caused by other problems in data collection and processing. They include:

1. **Overcoverage:** Inclusion of data from outside of the population.
2. **Undercoverage:** Sampling frame does not include elements in the population.
3. **Measurement error:** E.g. when respondents misunderstand a question, or find it difficult to answer.
4. **Processing error:** Mistakes in data coding.
5. **Non-response:** Failure to obtain complete data from all selected individuals.

After sampling, a review should be held of the exact process followed in sampling, rather than that intended, in order to study any effects that any divergences might have on subsequent analysis. A particular problem is that of *non-response*. Two major types of non-response exist: unit nonresponse (referring to lack of completion of any part of the survey) and item nonresponse (submission or participation in survey but failing to complete one or more components/questions of the survey).^{[10][11]} In survey sampling, many of the individuals identified as part of the sample may be unwilling to participate, not have the time to participate (opportunity cost),^[12] or survey administrators may not have been able to contact them. In this case, there is a risk of differences, between respondents and nonrespondents, leading

Random sampling by using lots is an old idea, mentioned several times in the Bible. In 1786 Pierre Simon Laplace estimated the population of France by using a sample, along with ratio estimator. He also computed probabilistic estimates of the error. These were not expressed as modern confidence intervals but as the sample size that would be needed to achieve a particular upper bound on the sampling error with probability 1000/1001. His estimates used Bayes' theorem with a uniform prior probability and it assumed his sample was random. The theory of small-sample statistics developed by William Sealy Gossett put the subject on a more rigorous basis in the 20th century. However, the importance of random sampling was not universally appreciated and in the USA the 1936 *Literary Digest* prediction of a Republican win in the presidential election went badly awry, due to severe bias [1]. More than two million people responded to the study with their names obtained through magazine subscription lists and telephone directories. It was not appreciated that these lists were

History

In many situations the sample fraction may be varied by stratum and data will have to be weighted to correctly represent the population. Thus for example, a simple random sample of individuals in the United Kingdom might include some in remote Scottish islands who would be inordinately expensive to sample. A cheaper method would be to use a stratified sample with urban and rural strata. The rural sample could be under-represented in the sample, but weighted up appropriately in the analysis to compensate. More generally, data should usually be weighted if the sample design does not give each individual an equal chance of being selected. For instance, when households have equal selection probabilities but one person is interviewed from within each household, this gives people from large households a smaller chance of being interviewed. This can be accounted for using survey weights. Similarly, households with more than one telephone line have a greater chance of being selected in a random digit dialing sample, and weights can adjust for this. Weights can also serve other purposes, such as helping to correct for non-response.

Survey weights

to biased estimates of population parameters. This is often addressed by improving survey design, offering incentives, and conducting follow-up studies which make a repeated attempt to contact the unresponsive and to characterize their similarities and differences with the rest of the frame.^[13] The effects can also be mitigated by weighting the data when population benchmarks are available or by imputing data based on answers to other questions. Nonresponse is particularly a problem in internet sampling. Reasons for this problem include improperly designed surveys,^[11] over-surveying (or survey fatigue),^{[14][15]} and the fact that potential participants hold multiple e-mail addresses, which they don't use anymore or don't check regularly. Web-based surveys also tend to demonstrate nonresponse bias; for example, studies have shown that females and those from a white/Caucasian background are more likely to respond than their counterparts.^[16]

heavily biased towards Republicans and the resulting sample, though very large, was deeply flawed.

References

- Ader, H. J., Mellenbergh, G. J., & Hand, D. J. (2008). Advising on research methods: A consultant's companion. Huiszen, The Netherlands: Johannes van Kessel Publishing.
- Bartlett, J. E., II, Kotlik, J. W., & Higgins, C. (2001). Organizational research: Determining appropriate sample size for survey research. *Information Technology, Learning, and Performance Journal*, 19(1) 43–50.
- Chambers, R. L., and Skinner, C. J (editors) (2003), *Analysis of Survey Data*, Wiley, ISBN 0-471-89987-9
- Cochran, William G. (1977). *Sampling Techniques* (Third ed.). Wiley. ISBN 0-471-16240-X.
- Deming, W. Edwards (1975) On probability as a basis for action, *The American Statistician*, 29(4), pp146–152.
- Deming, W. Edwards (1966). *Some Theory of Sampling*. Dover Publications. ISBN 0-486-64684-X. OCLC 166526.
- Gy, P (1992) *Sampling of Heterogeneous and Dynamic Material Systems: Theories of Heterogeneity, Sampling and Homogenizing* Wiley, ISBN 0-471-10949-5
- Kish, Leslie (1995) *Survey Sampling*, Wiley, ISBN 0-471-10949-5
- Korn, E I., and Graubard, B I (1999) *Analysis of Health Surveys*, Wiley, ISBN 0-471-13773-1
- Lohr, Sharon L. (1999). *Sampling: Design and Analysis*. Duxbury.
- Pedhazur, E., & Schmelkin, L. (1991). *Measurement design and analysis: An integrated approach*. New York: Psychology Press.
- Smdal, Carl-Erik, and Swensson, Bengt, and Wretman, Jan (1992). *Model Assisted Survey Sampling*. Springer-Verlag. ISBN 0-387-40620-4.
- Stuart, Alan (1962) *Basic Ideas of Scientific Sampling*, Hafner Publishing Company, New York
- Smith, T. M. F. (1984). "Present Position and Potential Developments: Some Personal Views: Sample surveys". *Journal of the Royal Statistical Society, Series A (General)* 147 (The 150th Anniversary of the Royal Statistical Society): pp. 208–221. <http://www.jstor.org/stable/2981677>. JSTOR 2981677
- Smith, T. M. F. (1993). "Populations and Selection: Limitations of Statistics (Presidential address)". *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 156 (2): pp. 144–166.
- <http://www.jstor.org/stable/2982726>. JSTOR 2982726 (Portrait of T. M. F. Smith on page 144)
- Smith, T. M. F. (2001). "Biometrika centenary: Sample surveys". *Biometrika* 88, (1): pp. 167–243. doi:10.1093/biomet/88.1.167. <http://biomet.oxfordjournals.org/cgi/content/abstract/88/1/167>.
- Smith, T. M. F. (2001). "Biometrika centenary: Sample surveys". *Biometrika* 88, (1): pp. 167–243. doi:10.1093/biomet/88.1.167.
- Titterton and D. R. Cox. *Biometrika: One Hundred Years*. Oxford University Press. pp. 165–194. ISBN 0-19-850993-6.

- Whittle, P. (May 1954). "Optimum preventative sampling". *Journal of the Operations Research Society of America* 2 (2): pp. 197-203.
<http://www.istor.org/stable/166605>.
- ASTM E105 Standard Practice for Probability Sampling Of Materials
- ASTM E122 Standard Practice for Calculating Sample Size to Estimate, With a Specified Tolerable Error, the Average for Characteristic of a Lot or Process
- ASTM E141 Standard Practice for Acceptance of Evidence Based on the Results of Probability Sampling
- ASTM E1402 Standard Terminology Relating to Sampling
- ASTM E1994 Standard Practice for Use of Process Oriented AOQL and LTPD Sampling Plans
- ASTM E2234 Standard Practice for Sampling a Stream of Product by Attributes Indexed by AQL

Survey Design Tips

Writing great questions is an art that like all arts requires a great amount of work, practice, and help from others. The following discussion is one that identifies some of the common pitfalls in survey design and helps in creating a great survey.

Avoid loaded or leading words or questions

Slight wording changes can produce great differences in results. Could, Should, Might all sound almost the same, but may produce a 20% difference in agreement to a question (The supreme court could.. should.. might.. have forced the breakup of Microsoft Corporation). Strong words that represent control or action, such as prohibit produces similar results (Do you believe that congress should prohibit insurance companies from raising rates?) Sometimes wording is just biased: You wouldn't want to go to Rudolph's Restaurant for the company's annual party would you?

Misplaced questions

Questions placed out of order or out of context should be avoided. In general, a funnel approach is advised. Broad and general questions at the beginning of the questionnaire as a warm-up. Then more specific questions, followed by more general easy to answer questions like demographics.

Mutually non-exclusive response categories

Multiple choice response categories should be mutually exclusive so that clear choices can be made. Non-exclusive answers frustrate the respondent and make interpretation difficult at best.

Nonspecific questions

Do you like orange juice? This is very unclear...do I like what? Taste, texture, nutritional content, Vitamin C, the current price, concentrate, fresh squeezed? Be specific in what you want to know about. Do you watch TV regularly? (what is regularly?).

Confusing or unfamiliar words

Asking about caloric content, bits, bytes, mbs, and other industry specific jargon and acronyms are confusing. Make sure your audience understands your language level, terminology and above all, what you are asking.

Non-directed questions give respondents excessive latitude

What suggestions do you have for improving tomato juice? The question is about taste, but the respondent may offer suggestions about texture, the type of can or bottle, mixing juices, or something related to use as a mixer or in recipes.

Forcing answers

Respondents may not want, or may not be able to provide the information requested. Privacy is an important issue to most people. Questions about income, occupation, finances, family life, personal hygiene and beliefs (personal, political, religious) can be too intrusive and rejected by the respondent.

Non-exhaustive listings
Do you have all of the options covered? If you are unsure, conduct a pretest using the "Other (please specify)" option. Then revise the question making sure that you cover at least 90% of the respondent answers.

Unbalanced listings
Unbalanced scales may be appropriate for some situations and biased in others. When measuring alcohol consumption patterns, One study used a quantity scale that made the heavy drinker appear in the middle of the scale with the polar ends reflecting no consumption and an impossible amount to consume. However, we expect all hospitals to offer good care and may use a scale of excellent, very good, good, fair. We do not expect poor care.

Double barreled questions
What is the fastest and most convenient Internet service for you? The fastest is certainly not the most economical. Two questions should be asked.

Dichotomous questions
Make sure answers are independent. For example the question "Do you think basketball players are being independent agents or as employees of their team?" Some believe that yes, they are both.

Long questions
Multiple choice questions are the longest and most complex. Free text answers are the shortest and easiest to answer. When you increase the length of questions and surveys, you decrease the chance of receiving a completed response.

Questions on future intentions.

Yogi Berra once said that making predictions is difficult, especially when they are about the future. Predictions are rarely accurate more than a few weeks or in some case months ahead.

Processing and interpreting data

Design a questionnaire or an experiment to capture the data needed to follow lines of enquiry and to test hypotheses, taking possible bias into account.

Questions for questionnaires

The two basic types of questions used in questionnaires are multiple-choice and open-ended (or free-answer). One disadvantage of some multiple-choice questions is that they may lead respondents to give answers which would not have been given if no responses had been suggested. The big disadvantage of open-ended questions is that the responses often have to be categorised before any statistical analysis can take place.

The following points about all questions used in questionnaires should be noted.

- Questions must be clear and unambiguous.
- Questions should be expressed in language appropriate for the target group.
- Questions should not require information which respondents are unlikely to possess or unwilling to reveal.
- Questions concerning measurements should state the units required.
- Leading questions should be not be used in order to avoid bias.

The following points concerning multiple-choice questions should also be noted.

- It should be made clear how and where the response should be recorded.
- The responses in the given list should be mutually exclusive.
- The response list should be exhaustive.
- The response list should not be such that the majority will give an 'Other' response.

Worked examples

Example 1

Sian wants to find out what people think about television programmes. She decides to carry out a survey using a questionnaire outside a supermarket on Saturday afternoon. Questions 23 and 24 on her questionnaire are given below.

23 Most people think that there is too much sport on television. Do you agree?

YES NO

24 What is your favourite type of TV programme?

- Criticise each of the above questions.
- Write better versions of Q23 and Q24.
- Give one reason why her survey is likely to be biased.

Answer	
(a)	e.g. Q23 is a leading question. or It is not clear how the answer should be recorded.
(b)	e.g. Q24 is ambiguous because the meaning of the word 'type' is unclear. or The open-ended nature of the question would result in difficulties in categorising the answers.
(b)	Q23 What do you think about the amount of sport on television? Place a tick in one of the boxes below.
	Too little About right Too much
Q24	Indicate your favourite type of television programme by placing a tick in one of the boxes below.
	Soap Film/ Drama Comedy News/ Current Affairs Sport Wildlife Other
(c)	e.g. People who like sport on TV are likely to be under-represented because many will be watching the Saturday afternoon sports programmes or attending matches. or There are too many questions (twenty four at least) on the questionnaire only very patient people, who may have different preferences, will answer.

Handling data

Example 2

Mr Smith, who teaches PE and Health Education, wishes to test the hypothesis 'smoke more than boys'. He plans to give the following questionnaire to pupils in his classes.

Questionnaire on smoking	
1 What is your name and form?	
2 What is your sex?	Male / Female
3 How many cigarettes do you smoke per day? Tick one of the boxes below	None 1 - 5 5 - 10 10 - 15 15 - 20

- Give a reason why Mr Smith's results may be biased?
- State three faults in the questionnaire.
- Give an improved version of the questionnaire.

Answer	
(a) e.g. Mr Smith's PE classes are likely to be single sex, so the total number of boys in all his classes is likely to be greater than the total number of girls or The boys and girls could be unmatched with regard to age.	
(b) The first question may discourage pupils from giving truthful answers. The responses listed in Q3 are not mutually exclusive. The responses listed in Q3 are not exhaustive.	
(c) Questionnaire on smoking	
1 What was your age at your last birthday?	years
2 What is your sex?	Male / Female
3 How many cigarettes do you smoke per day? Tick one of the boxes below	None 1 - 5 6 - 10 11 - 15 Over 15

THE COLLECTION AND DISPLAY OF STATISTICS

ESSENTIAL SKILLS

This first section about statistics covers the early skills. Very little is assumed.

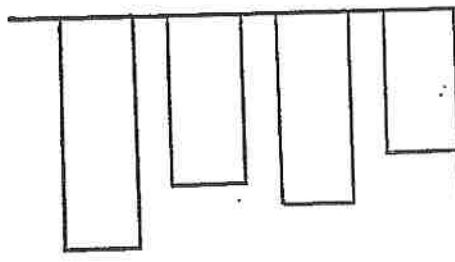
BASIC FACTS

Data is collected in this way:

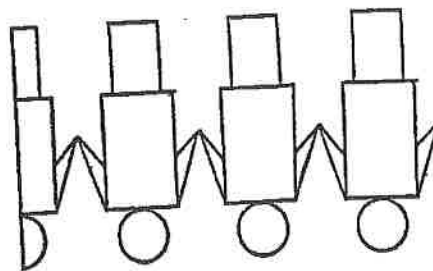
Group	Tally	f
1 - 5	<del> </del>	7
6 - 10		5
11 - 15		5
16 - 20		5
21 - 25		5
26 - 30		5
31 - 35		5
36 - 40		5
41 - 45		5
46 - 50		5
51 - 55		5
56 - 60		5
61 - 65		5
66 - 70		5
71 - 75		5
76 - 80		5
81 - 85		5
86 - 90		5
91 - 95		5
96 - 100		5
etc.		

Several simple diagrams can be used to display data, including:

(a) bar charts



(b) pictograms



Diagrams can mislead. Examples will be given.

In statistics we handle numbers for various reasons. However, whatever our intentions we must collect this information first.

Examples

- Here are the results of a spelling test which was taken by 25 pupils.
The maximum mark was 5.

1	3	2	3	1
3	2	4	4	3
2	3	3	2	3
3	2	4	2	3
4	5	3	4	2
3	3	2	1	1

This jumble of numbers needs to be put in order. A table (as below), called a frequency table, is a way of sorting the numbers. To make the final adding easy, use each fifth stroke to cross the other four, like this ~~||||~~. Here is the frequency table of the test results above:

Score	Tally	Frequency (f)
1		2
2	<del> </del>	7
3	<del> </del>	10
4	<del> </del>	5
5	/	1

Total 25 (a useful check)

If the marks had been more spread out, we may have wanted to put them into groups.

- Here are the ages of 30 people who attended a pop concert:

16	8	18	17
6	21	28	12
11	18	18	23
16	13	19	22
17	21	14	24
13	17	16	19

The ages range from 6 to 28. Groups of 1 - 5, 6 - 10, 11 - 15, etc. have been chosen. The table looks like this:

Age	Tally	f
1 - 5		3
6 - 10	<del> </del>	7
11 - 15	<del> </del>	14
16 - 20	<del> </del>	5
21 - 25	<del> </del>	1
26 - 30	/	1

Total 30 (checking again)

EXERCISE

- 40 people were asked to choose their favourite comedy T.V. programme from the following:

A Alas Smith & Jones
 B Yes Prime Minister
 C The Benny Hill Show
 D The Two Ronnies
 E The Lenny Henry Show

Their responses were:

A C C A
 C B D D
 E B B C
 B E B B
 E E E E
 A A A E
 C B E B
 D B A D
 B A B D
 C B B C

Draw up a table of their choices.

- Here are the mathematics test results of 48 G.C.S.E. students. The

maximum mark was 50.
 Draw up a table of these results using groups 1 - 10, 11 - 20, 21 - 30, etc.

12	23	31	46	42	46	29
27	31	26	38	7	28	43
19	27	36	31	36	11	36
19	24	31	34	22	27	5
19	28	31	41	28	19	31
32	41	40	9	36	28	27

Answers to the Exercise

- Choice

A	8
B	14
C	6
D	4
E	8

Tally

Total 40

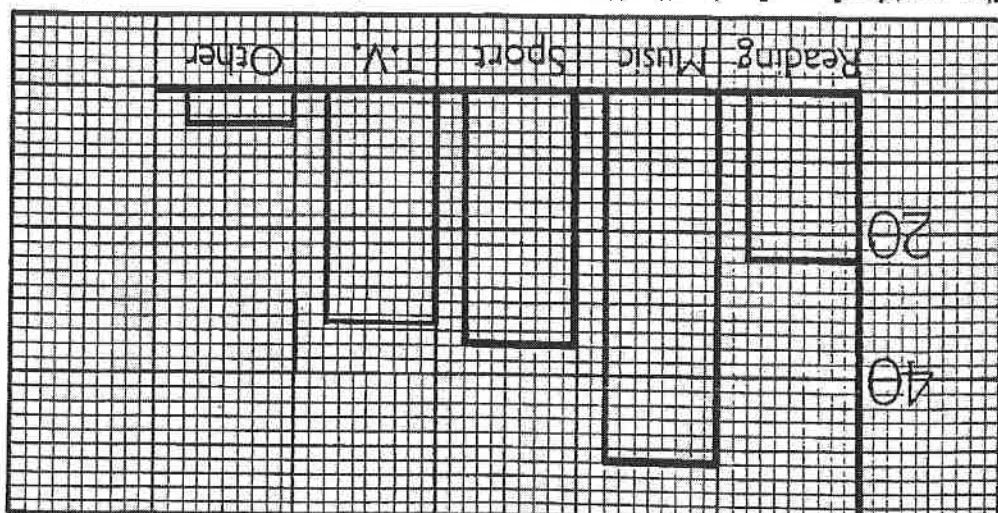
- Group

1 - 10	3
11 - 20	6
21 - 30	15
31 - 40	15
41 - 50	9

Tally

Total 48

The bar graph below illustrates the favourite interests of some adults:



The vertical scale is the important one. Check that you can 'read' a small square. In this case, as 1 cm represents 10 adults, then one small square (2 mm) represents two adults.

3. Use this bar graph to answer these questions:

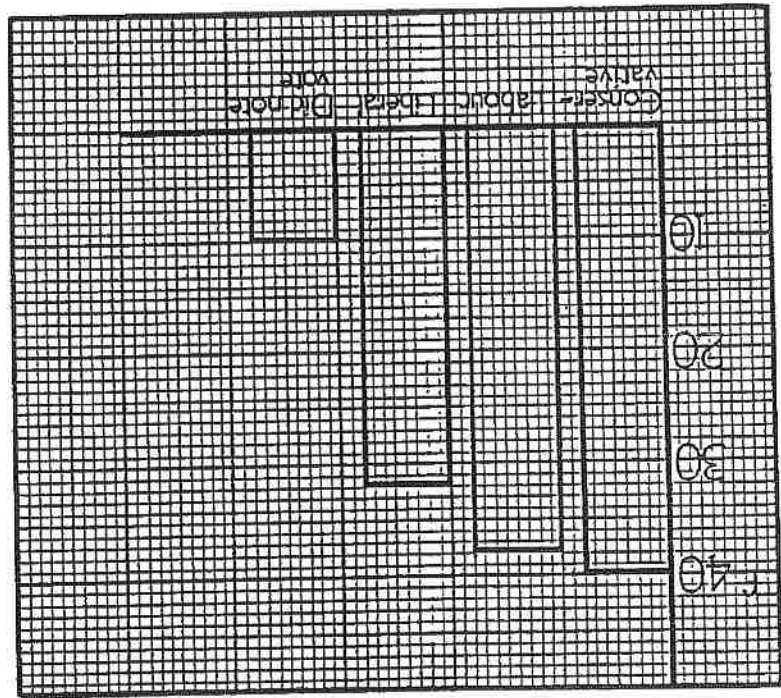
- What is the most popular hobby?
- How many chose sport?
- How many more adults chose T.V. than reading?
- What is the total number of adults represented in the diagram?

Solution:

- music (the highest column)
- 36 (33 would be a common error)
- $33 - 24 = 9$
- $24 + 52 + 36 + 33 + 5 = 150$

EXERCISE

1. A mock election was held in a school and the way in which the sixth-formers voted was noted on this graph.
Use this bar graph to answer the questions which follow.



- (a) How many people are represented by one small square (2 mm) on the vertical axes?
- (b) How many voted Liberal?
- (c) How many more people voted Labour than Liberal?
- (d) How many sixth-formers voted?

2. Draw a bar graph to illustrate this information:

The monthly T.V. viewing (in hours) of an average adult is:

Programme	Hours
serials	34 hours
sport	21 hours
documentaries and news	17 hours
comedy programmes	13 hours
others	9 hours

3. The following vehicles passed a school in 30 minutes:

Vehicle	Count
cars	48
buses	6
bicycles	7
lorries and vans	29
motorcycles	16

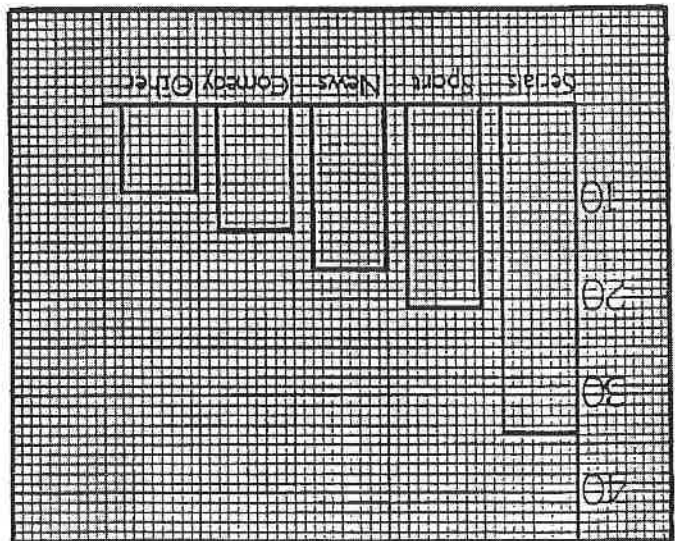
Illustrate this information with a bar chart.

Answers to the Exercise

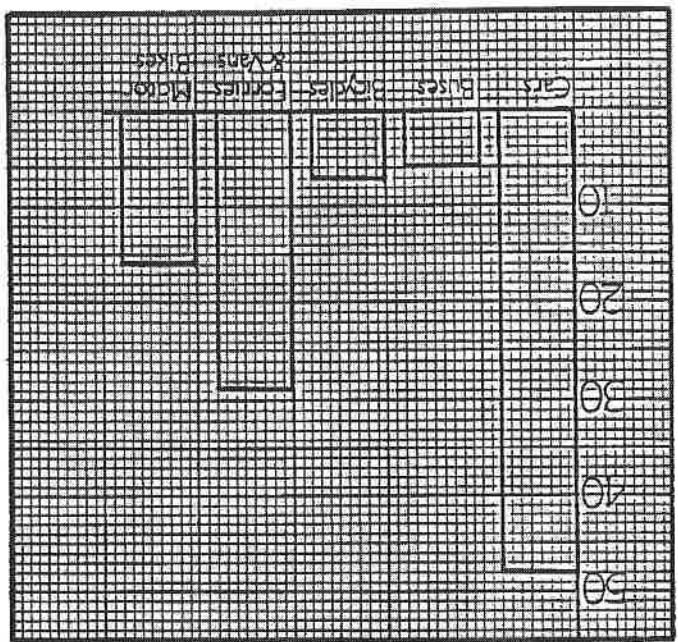
1. (a) 1 (b) 32 (c) 6 (d) 120

Your graphs should look like this:

2.

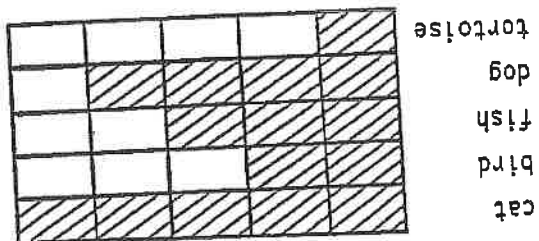


3.

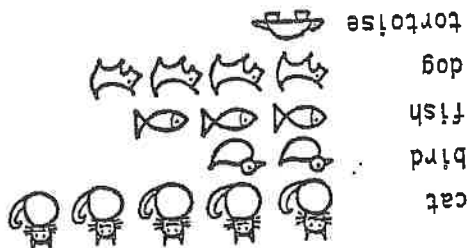


Pictograms

If a bar chart is to be shown on television, or is to be printed in a newspaper, it is often dressed up so that each column (which is simply a rectangle) becomes a shape linked to the information given.
e.g. if the size of families is being depicted,
3 units become 3 children
if it is the sale of milk, 5 units become 5 milk bottles, etc.
the pets of some young children were first drawn as a block graph:



and then the same information was displayed as a pictogram:

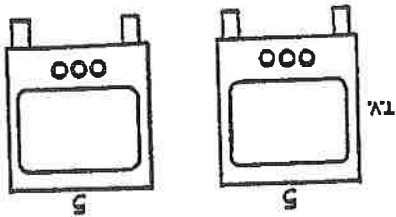


EXERCISE

- Design suitable symbols to illustrate:
 - the various forms of transport used by people to get to work,
 - the types of houses inhabited by people in a town, e.g. flats, detached houses, etc.
 - the favourite hobbies of the adults in the earlier question, i.e. music, sport, T.V. and reading.
- Draw a pictogram, using your symbols, for question 1.(c) if the figures are:

music	25
sport	30
T.V.	12
reading	16

Your attempt at Question 2 above will bring home to you the main problems with pictograms. For example, if you let 1 T.V. set represent 5 adults, you will have got the problem of showing 12 adults. Two and a bit T.V. sets is too vague.



Now what?

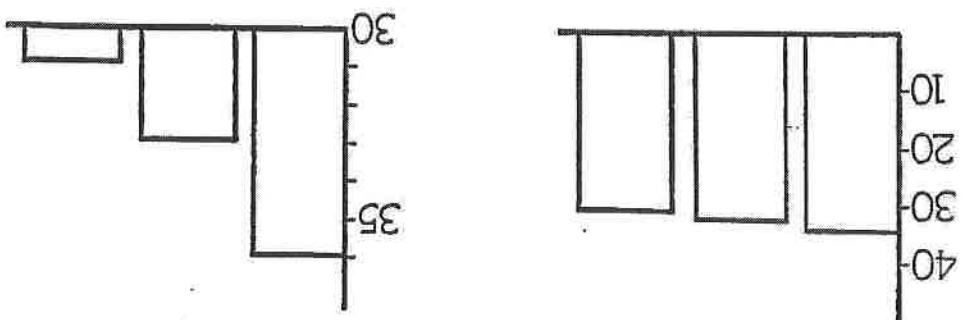
A good pictogram is supposed to give numerical facts as well as being interesting.

Misleading Graphs

Here are a few mistakes which are made when drawing various graphs. On occasions these mistakes are made deliberately. Don't make them yourself. Understand the weaknesses. You may well have to criticise poor graphs in the examination.

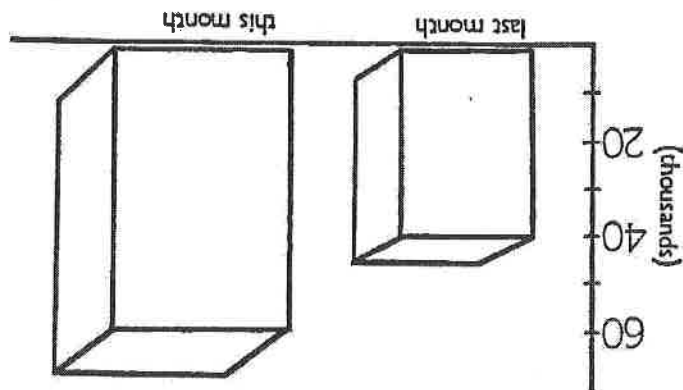
1. The two bar graphs below illustrate the sales figures claimed by a company selling washing-up liquid. The figures are:

their sales 36%
the sales of their nearest rival 33%
the rest 31%



By starting the vertical scale at 30 in the second graph, the ratio of sales is hopelessly distorted, although if you look closely, the facts are the same.

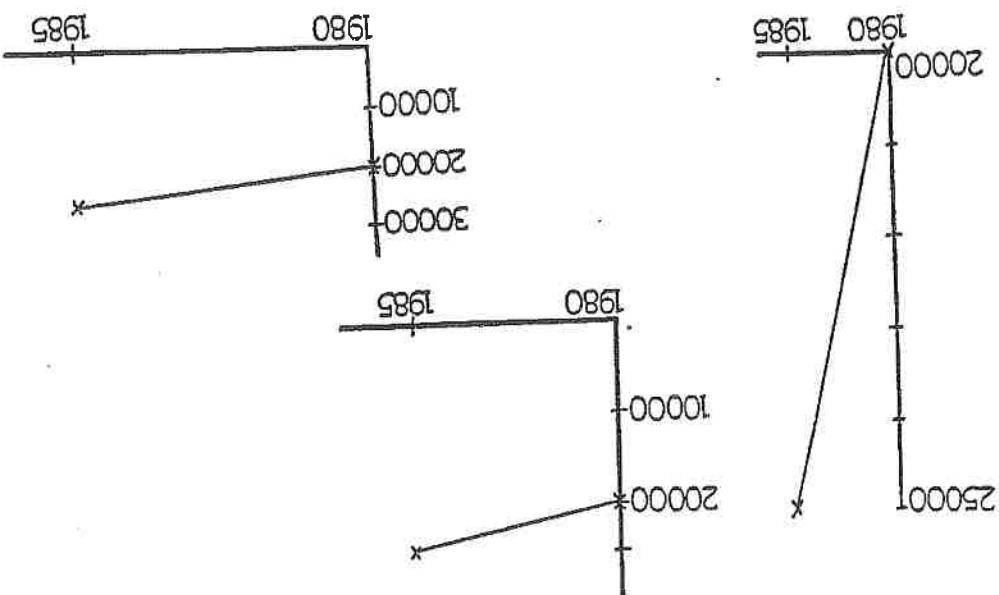
2. A soap powder company's sales increased from 40 000 packets per month to 60 000 packets per month. This pictogram illustrates this fact:



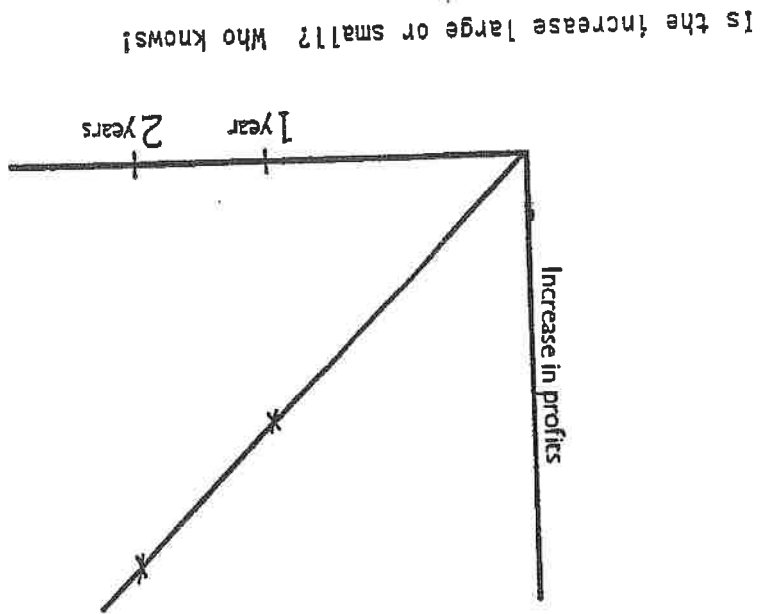
The vertical scale is correct, but, as the boxes' lengths and breadths are also to scale, at first glance the wrong impression is given.

3. The choice of scale can mislead as well. These three graphs all

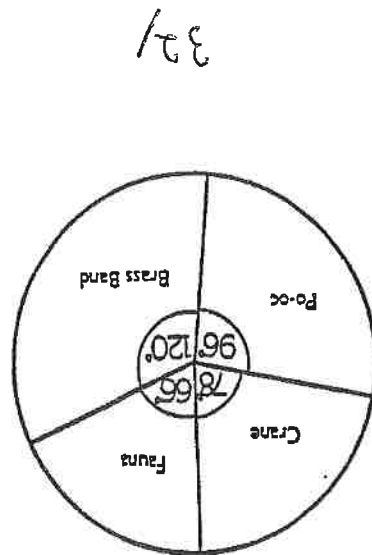
illustrate the same information.
In 1980 a bank had 20 000 customers with current accounts. In 1985, it had 25 000 customers with current accounts.



4. The graphs speak for themselves.
Finally, the graphs without scales. They tell us nothing.



Is the increase large or small? Who knows!



The pie chart can now be drawn. Label it carefully.
(Note : $66 + 78 + 96 + 96 + 120 = 360^\circ$; a useful check)

Fauna	$11 \times 6 = 66^\circ$
Crane	$13 \times 6 = 78^\circ$
Po-oc	$16 \times 6 = 96^\circ$
Brass Band	$20 \times 6 = 120^\circ$

1 customer is represented by $360 \div 60 = 6^\circ$

Solution Show your calculations.

Draw a pie chart to illustrate this table.

Fauna	11
Crane	13
Po-oc	16
Brass Band	20

1. 60 customers were asked to name their favourite margarine. The results were as follows:

Examples

Pie charts are excellent for comparing figures. Each quantity is represented by a sector of a circle. As the angles at a point add to 360° , this number figures a great deal.

PIE CHARTS

A pie chart presents information by cutting a circle into sectors. 360° is the important angle. Histograms are similar to bar charts, but care is needed with axes and scales.

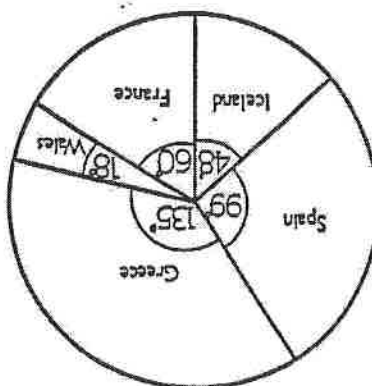
BASIC FACTS

- (a) Care with a protractor and a knowledge of angles at a point
(b) Ease with the reading and construction of bar charts (See section 23)

ESSENTIAL SKILLS

HISTOGRAMS AND PIE CHARTS

2. 120 people were asked to choose in which of five countries they would like to spend their next summer holiday. This pie chart illustrates their answers:



- (a) Which country was most popular?
 Clearly, Greece.
 (b) Which country was least popular?
 Wales, I'm afraid.
 (c) How many people chose Spain?

Pie charts do not provide the actual figures easily. However, if 120 people are represented by 360°, 1 person is represented by 3°. So, the number of people choosing Spain was $\frac{48}{3}$ or 33.

Histograms

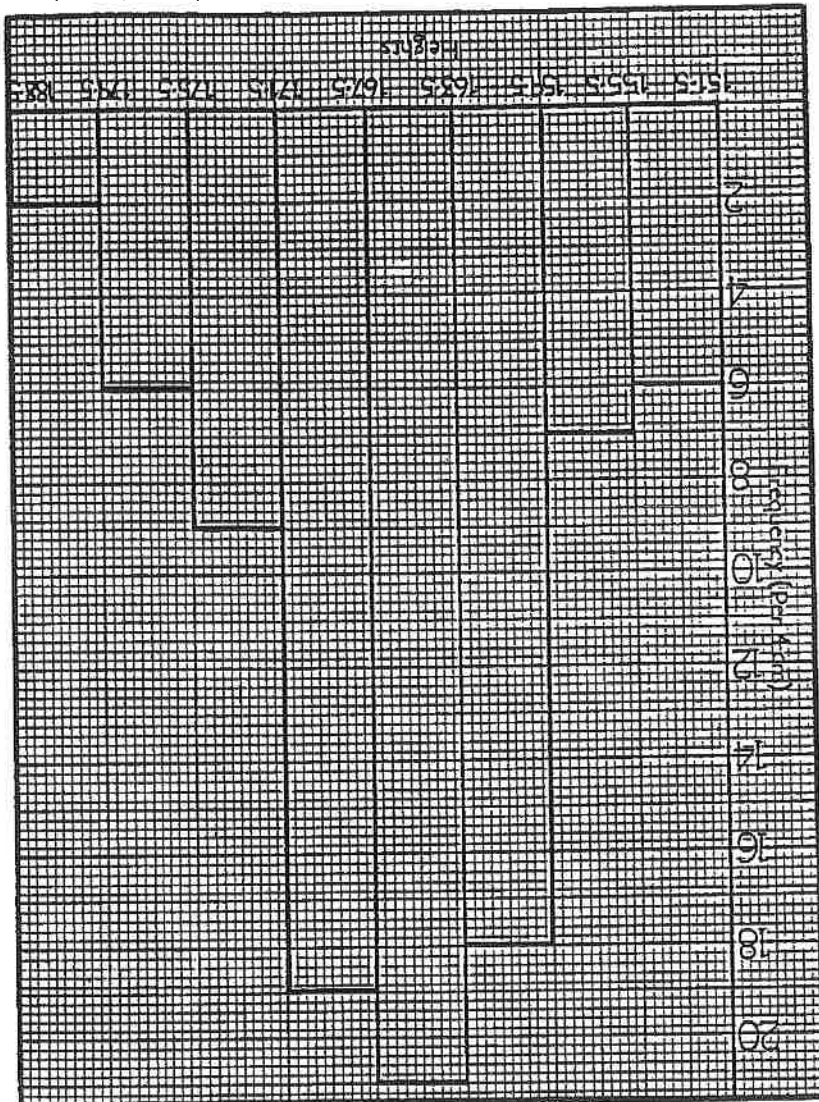
Histograms resemble bar charts. However, it is the area of the bar in a histogram which represents the frequency. Frequency densities are usually calculated. Fortunately the bars of the histograms that we are asked to consider are of equal widths (class intervals) and it is acceptable (more or less) that the heights of our bars can be taken from the table provided.

Examples

3. The heights of 88 girls in a sixth form were measured in cm. The results are below:

Class Interval	Frequency
151.5 - 155.5	6
155.5 - 159.5	7
159.5 - 163.5	18
163.5 - 167.5	21
167.5 - 171.5	19
171.5 - 175.5	9
175.5 - 179.5	6
179.5 - 183.5	2

Draw a histogram to illustrate this data.

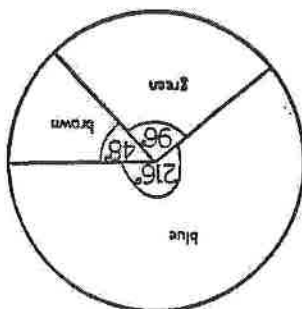


Notice that:

- (a) the horizontal scale starts at the smallest class boundary and finishes at the highest
- (b) there are no gaps
- (c) by using 4 cm as the unit of height, the frequency table was plotted directly

EXERCISE

- This pie chart illustrates the eye colour of 30 children in a class.



- Use your pie chart to find:
 - the most common eye colour
 - the least common eye colour
 - what angle represented one child?
 - How many children had green eyes?

- A student spent 24 hours revising for her G.C.S.E mock examinations. The time was spent on four subjects in this way:

Mathematics	10 hours
Science	8 hours
History	5 hours
French	1 hour

Draw a pie chart to illustrate this information

- 80 boys ran 400 m. Here are their times in seconds:

Time	
59.5 - 64.5	7
64.5 - 69.5	10
69.5 - 74.5	36
74.5 - 79.5	17
79.5 - 84.5	10

Draw a histogram to illustrate these figures.

- The weights of 100 newborn babies are provided in the table below. Draw a histogram.

Weights (in g)	
2000 - 2500	6
2500 - 3000	19
3000 - 3500	38
3500 - 4000	28
4000 - 4500	9

- The lives of 25 light bulbs were measured to the nearest hour. They were

	76	83	96	88	75
	71	83	104	96	92
	87	84	93	86	91
	72	86	107	93	87
	88	93	86	94	109

Copy and complete this frequency distribution:

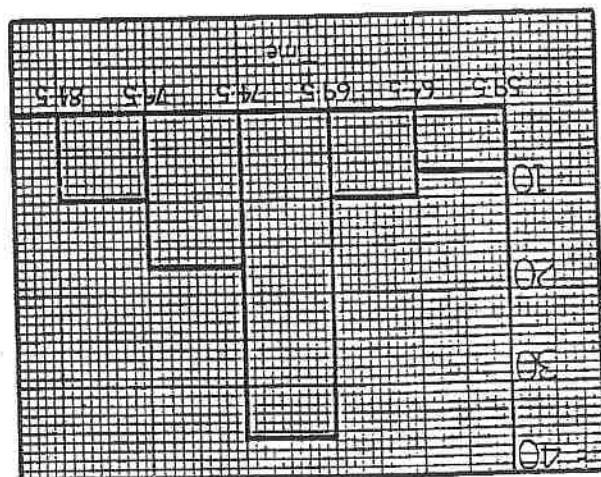
Life (nearest hour)	Tally	f
71 - 80	111	4
81 - 90	1111	10
91 - 100	1111	10
101 - 110	111	4

Draw a histogram to illustrate your frequency distribution.

- The answers are:
1. (a) blue (ii) brown
 - (b) 12°
 - (c) 8

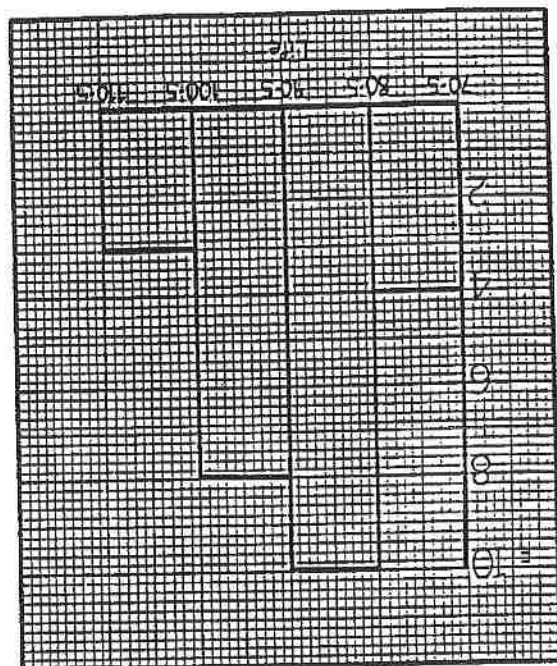
2. The angles were 150° , 120° , 75° and 15°

3.

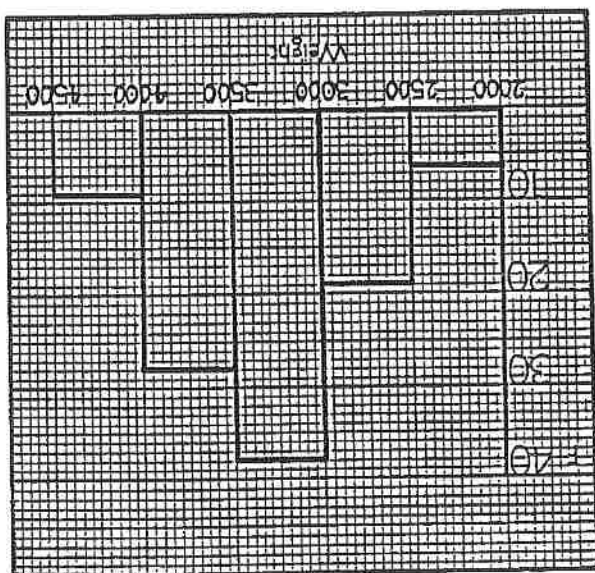


5.

Life		Tally		f	
71 - 80	1111		111	10	4
81 - 90			111	10	8
91 - 100			111	10	8
101 - 110			111	10	3



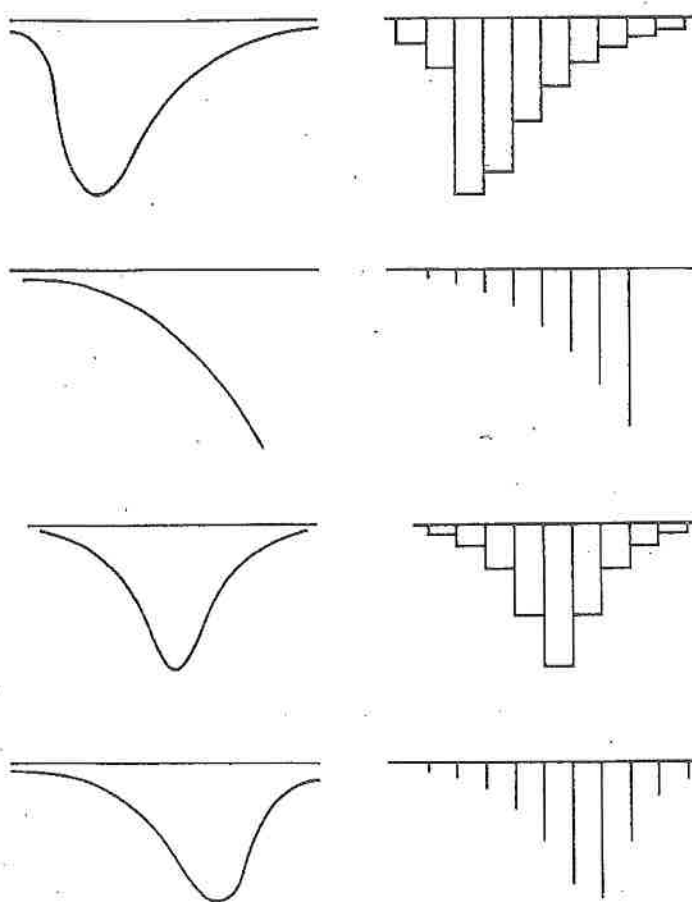
4.



SHAPES OF DISTRIBUTIONS

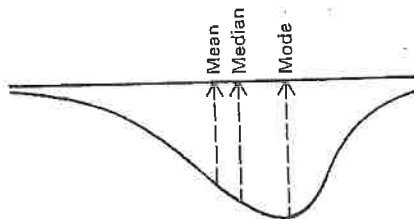
THE SHAPE OF A DISTRIBUTION

If distributions represented by a vertical line graph or a histogram are illustrated using a frequency curve, it is easier to see the general 'shape' of the distribution. For example:



Positively Skewed Distribution

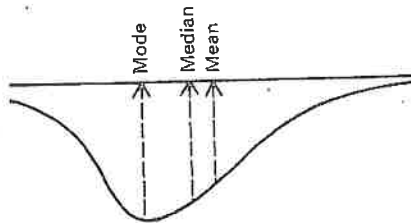
In a positively skewed distribution, the mean is pulled in a positive direction (→). The frequency curve would be:



This type of distribution occurs when considering, for example, the number of children per family, the age at which women marry, the distribution of wages in a firm.

NEGATIVELY SKEWED DISTRIBUTIONS

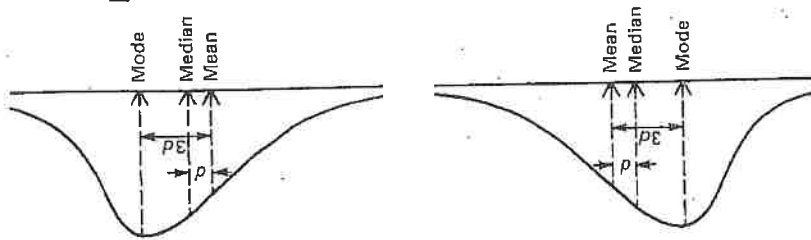
In a negatively skewed distribution, the mean is pulled in a negative direction (←). The frequency curve would be:



This type of distribution can occur when considering, for example reaction times for an experiment, daily maximum temperatures for a month in summer.

NOTE: the median generally lies between the mode and the mean and the following relation is satisfied

$$\text{mean} - \text{mode} \approx 3[\text{mean} - \text{median}]$$



$$\text{Mean} - \text{Mode} \approx 3[\text{Mean} - \text{Median}]$$

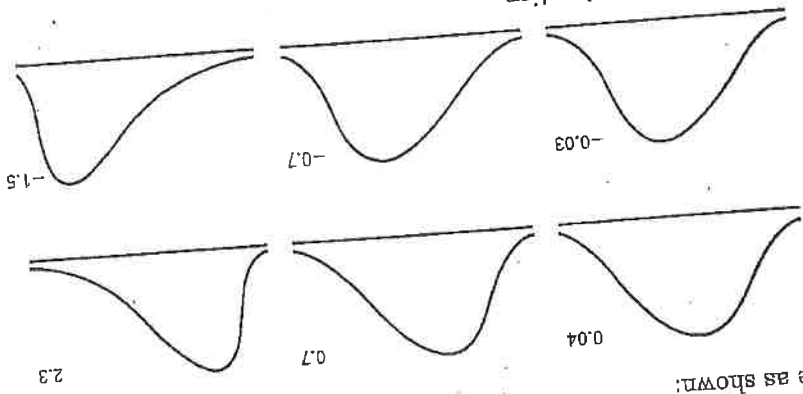
We need to know two things to describe the 'skewness' of a distribution. One is the direction of the skew, positive or negative, and the other is a measure of the degree of skewness, i.e. how far the distribution is pulled in one direction. One measure of skewness is given by

Pearson's coefficient of skewness

$$\text{Pearson's coefficient of skewness} = \frac{\text{mean} - \text{mode}}{\text{standard deviation}}$$

- If mean > mode, the skew is positive.
- If mean < mode, the skew is negative.
- If mean = mode, the skew is zero and the distribution is symmetrical.

Generally skewness can take any values between 3 and -3. For example, the measure of skewness for these distributions might be as shown:



Using the approximation
 $\text{mean} - \text{mode} = 3(\text{mean} - \text{median})$

the formula for skewness becomes

$$\text{Pearson's coefficient of skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

Example

A farmer records the number of eggs he collects from his hens over a 150-day period. The frequency distribution of the eggs collected is shown below.

Number of eggs	Frequency
15	1
16	2
17	4
18	6
19	9
20	13
21	18
22	22
23	24
24	30
25	10

What is the modal number of eggs collected per day?

is 2.18 (2 d.p.).

Therefore the mean is 21.94 and the standard deviation

$$= 2.179 \dots$$

$$s = \sqrt{4.749 \dots}$$

$$= 4.749 \dots$$

$$= \frac{72917}{150} - (21.94)^2$$

$$s^2 = \frac{\sum fx^2}{\sum f} - \bar{x}^2$$

$$= 21.94$$

$$= \frac{150}{3291}$$

$$\bar{x} = \frac{\sum fx}{\sum f}$$

(b) To find the mean and standard deviation, use

The median is 22.5 eggs.

value is 22.5.

So the 75th value is 22 and the 76th value is 23. Therefore the 75.5th

Number of eggs	...	53	75	110
Cumulative frequency	...	...	...	...

Part of the cumulative frequency table indicates:

i.e. the 75.5th value.

(a) $n = 150$, so the median is the $\frac{1}{2}(150 + 1)$ th value,

The modal number of eggs is 23.

Solution 2.

Calculate

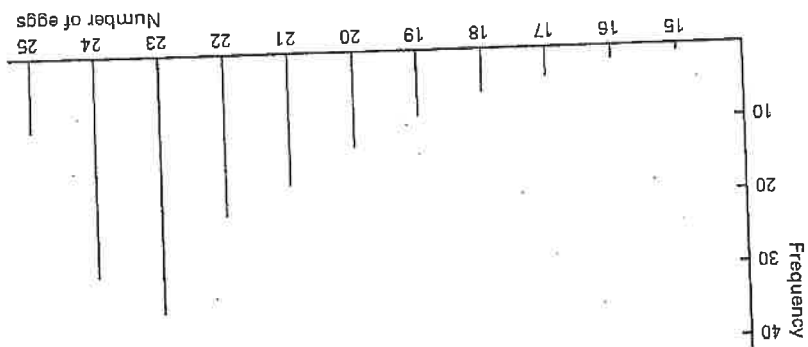
(a) the median number of eggs collected per day,

(b) the mean and standard deviation of the number of eggs collected per day.

of the diagram.

Draw a line diagram to illustrate the data and comment on the shape

Line graph to show number of eggs collected



We can see from the line diagram that the distribution is negatively skewed.
From the results obtained above we could calculate the skewness, where

$$\text{skewness} = \frac{\text{mean} - \text{mode}}{\text{standard deviation}} = \frac{21.94 - 23}{2.179...} = -0.486$$

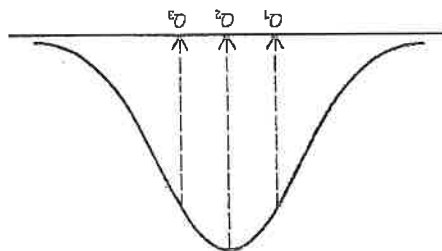
This also indicates that the distribution is negatively skewed.

QUARTILE COEFFICIENT OF SKEWNESS

Another measure of skewness is defined in terms of the quartiles, where q_1 is the lower quartile, q_2 is the median and q_3 is the upper quartile.

$$\text{Quartile coefficient of skewness} = \frac{(q_3 - q_1) - 2(q_2 - q_1)}{q_3 - q_1}$$

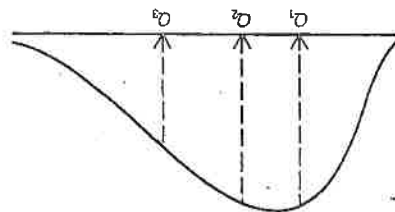
Normal distribution



$$q_3 - q_2 = q_2 - q_1$$

Quartile skewness = 0

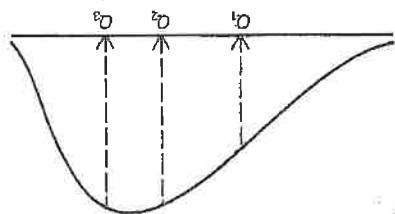
Positively skewed distribution



$$q_3 - q_2 > q_2 - q_1$$

Quartile skewness > 0

Negatively skewed distribution



$$q_3 - q_2 < q_2 - q_1$$

Quartile skewness < 0

31 students tried to estimate the length of a line. The line was actually 60 mm long. These are their results, in mm.

61 70 46 44 26 23 30 83 52 44 38
37 49 59 58 63 31 29 37 48 76 61
46 31 38 41 49 52 56 75 61

Find the median and the quartiles of this distribution. Use the quartiles to estimate the skewness. Draw a histogram using class intervals 20-, 30-, ..., 80-90.

We arrange the results in order. There are 31 results, so the median is the $\frac{1}{2}(31+1)$ th value, i.e. the 16th value.

To find the quartiles, since n is odd (see page 64)

$$Q_1 = \frac{1}{4}(31+1)\text{th value} = 8\text{th value}$$

$$Q_3 = \frac{3}{4}(31+1)\text{th value} = 24\text{th value}$$

23 26 29 30 31 31 37 (37) 38 38 41 44 44 46 46
(48) 49 49 52 52 56 58 59 (61) 61 61 63 70 75 76 83

median $Q_2 = 48$, $Q_1 = 37$, $Q_3 = 61$.

$$Q_3 - Q_2 = 61 - 48 = 13$$

$$Q_2 - Q_1 = 48 - 37 = 11$$

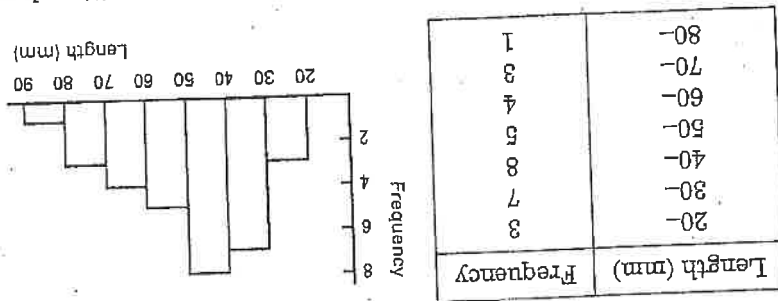
Since $Q_3 - Q_2 > Q_2 - Q_1$ the distribution is positively skewed.

$$\text{Quartile coefficient of skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

$$= \frac{13 - 11}{61 - 37}$$

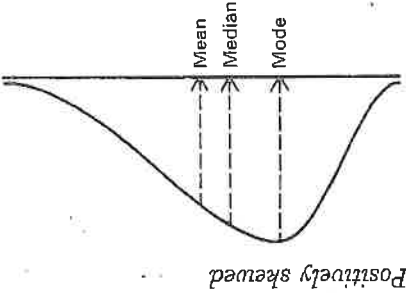
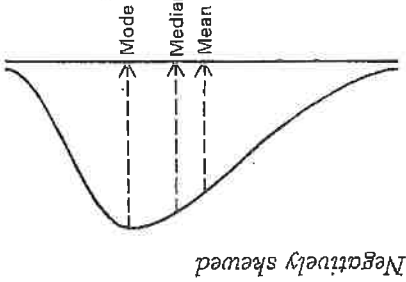
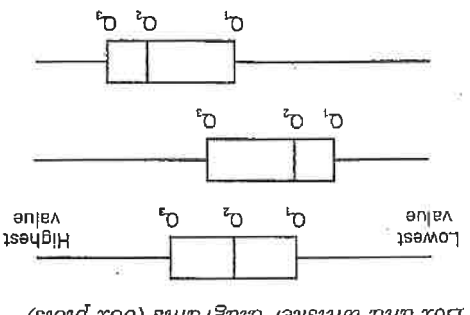
$$= 0.083 \dots$$

The frequency distribution is as shown:



The diagram confirms the positive skew.

SUMMARY

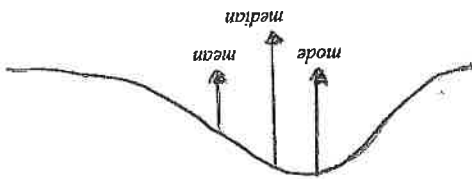
<i>Positively skewed</i>  mode < median < mean skewness > 0	<i>Negatively skewed</i>  mean < median < mode skewness < 0	<i>Approximate relationship between the three averages</i> mean - mode ≈ 3 (mean - median) <i>Skewness</i> Pearson's coefficient of skewness = $\frac{\text{mean} - \text{mode}}{\text{standard deviation}}$ $\approx \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$ Quartile coefficient of skewness = $\frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$	<i>Box and whisker diagrams (box plots)</i>  Symmetrical distribution Positively skewed distribution Negatively skewed distribution <i>Outliers</i> Take points lying more than 1.5 times the interquartile range above Q_3 or below Q_1 as 'outliers'.
--	--	--	---

SKewed DISTRIBUTIONS

The same ideas apply for a continuous distribution, and a little bit of thought should show that for positive skew $mean > median > mode$.



+ve skew



Negative skew

If a symmetrical box plot is stretched in the direction of the negative x-axis, then the resulting distribution has negative skew.

Q_1 Q_2 Q_3

PULL

Q_1 Q_2 Q_3

For negative skew the diagram shows that $Q_3 - Q_2 < Q_2 - Q_1$

The same ideas apply for a continuous distribution, and a little bit of thought should show that for negative skew $mean < median < mode$.



5 Measures of spread Range & interquartile range

Range

The *range* is found by subtracting the smallest value from the largest value.

Interquartile range

The *interquartile range* is found by subtracting the lower quartile from the upper quartile, so $I.Q.R. = Q_3 - Q_1$.

Variance and standard deviation

Variance is the square of the standard deviation.

$$s_{x^2}^2 = \frac{1}{N} \sum (x_i - \bar{x})^2 f_i, \text{ or}$$

$$s_{x^2}^2 = \frac{1}{N} \sum x_i^2 f_i - \bar{x}^2.$$

When finding the variance, it is nearly always easier to use the second formula.

Variance and standard deviation measure the *spread* of the distribution.

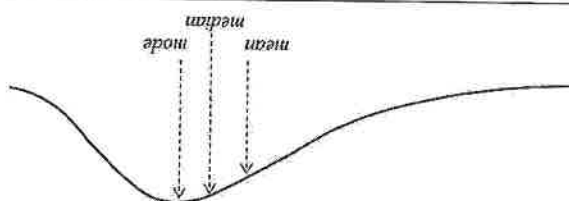
Proof of the alternative formula for variance

$$s_{x^2}^2 = \frac{1}{N} \sum (x_i - \bar{x})^2 f_i = \frac{1}{N} \sum (x_i^2 - 2x_i \bar{x} + \bar{x}^2) f_i$$

$$= \frac{1}{N} \sum x_i^2 f_i - \frac{1}{N} \sum 2x_i \bar{x} f_i + \frac{1}{N} \sum \bar{x}^2 f_i$$

$$= \frac{1}{N} \sum x_i^2 f_i - \frac{2\bar{x}}{N} \sum x_i f_i + \frac{\bar{x}^2}{N} \sum f_i$$

since $\bar{x} = \frac{1}{N} \sum x f$ and $N = \sum f$



177

$$s_t^2 = \frac{1}{N} \sum t_i^2 f_i - t^2 = \frac{141}{117} - \left(\frac{1}{39}\right)^2 = 1.204470743$$

and the variance of t is $s_t^2 = \sqrt{1.204470743} = 1.097483824 = 1.10$ to 3 s.f.

To find $\bar{x}$, using $t_i = \frac{x_i - 210}{5}$, $\Rightarrow \bar{x} = 5\bar{t} + 210 = 215.5$ to 1 D.P.

To find the standard deviation of x

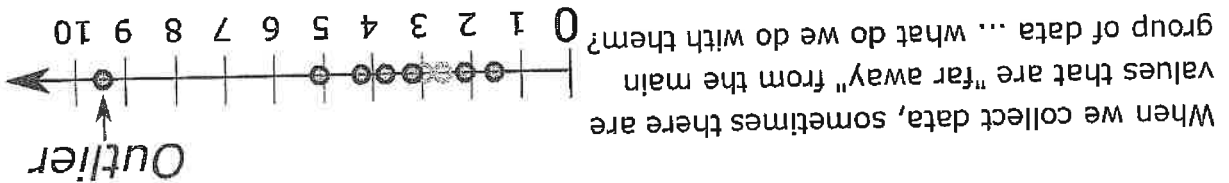
$$s_x = 5s_t = 5 \times 1.0974838... = 5.49$$
 to 3 s.f.

We would need to multiply the variance by $s^2 = 25$

$$\Rightarrow s_x^2 = 25s_t^2 = 25 \times 1.204470743 = 30.1$$
 to 3 s.f.

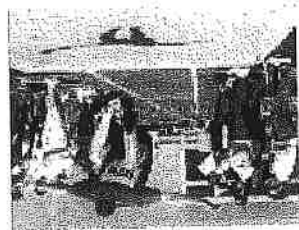
Outliers

"Outliers" are values that "lie outside" the other values.



Example: Long Jump

A new coach has been working with the Long Jump team this month, and the athletes' performance has changed.



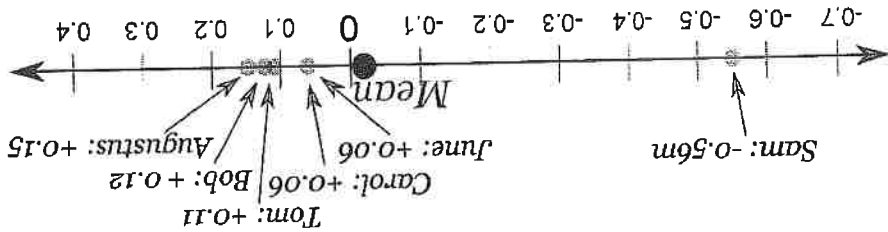
Augustus can now jump 0.15m further, June and Carol can jump 0.06m further.

Here are all the results:

- Augustus: +0.15m
- Tom: +0.11m
- June: +0.06m
- Carol: +0.06m
- Bob: + 0.12m
- Sam: -0.56m

Oh no! Sam got worse.

Here are the results on the number line:



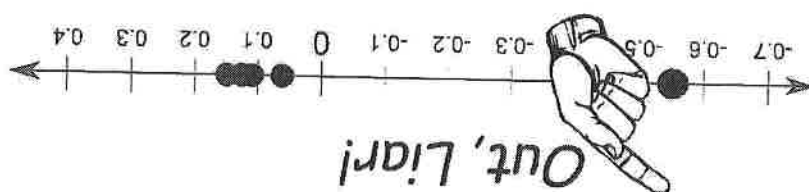
The **mean** is:

$$(0.15 + 0.11 + 0.06 + 0.06 + 0.12 - 0.56) / 6 = -0.06 / 6 = -0.01m$$

So, on average the performance went **DOWN**.

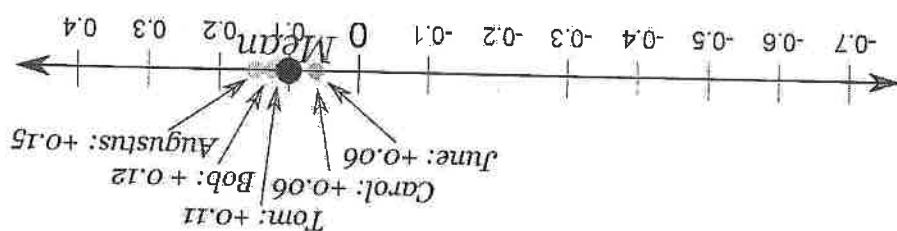
The coach is obviously useless ... right?

Sam's result is an "Outlier" ... what if we remove Sam's result?



Example: Long Jump (continued)

Let us try the results WITHOUT Sam:



$$\text{Mean} = (0.15 + 0.11 + 0.06 + 0.06 + 0.12) / 5 = 0.1m$$

Hey, the coach looks much better now!

But is that fair? Can we just get rid of values we don't like?

What To Do?

You need to think "why is that value over there?"

It may be quite normal to have high or low values

- People can be short or tall
- Some days there is no rain, other days there can be a downpour
- Athletes can perform better or worse on different days

Or there may be an unusual reason for extreme data

Example: Long Jump (continued)

We find out that Sam was feeling sick that day. Not the coach's fault at all.

So it is a good idea **in this case** to remove Sam's result.

When we remove outliers we are **changing the data**, it is no longer "pure", so we shouldn't just get rid of the outliers without a good reason!

And when we do get rid of them, we should **explain what we are doing and why**.

Mean, Median and Mode

We saw how outliers affect the mean, but what about the median or mode?

Example: Long Jump (continued)

The median ("middle" value):

- including Sam is: 0.085
- without Sam is: 0.11 (went up a little)

The mode (the most common value):

- including Sam is: 0.06
- without Sam is: 0.06 (stayed the same)

The mode and median didn't change very much.

Outliers

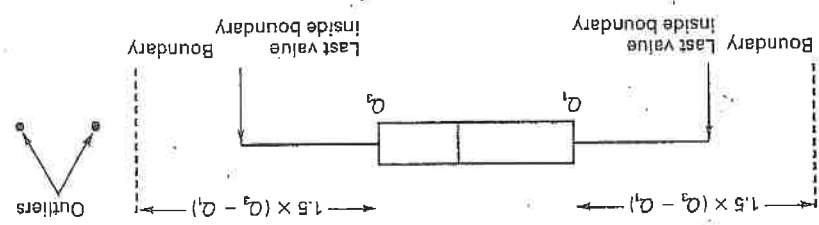
They also stayed around where most of the data is.

So it seems that outliers have the biggest effect on the mean, and not so much on the median or mode.

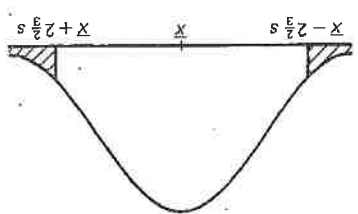
Hint: calculate the median and mode when you have outliers.

A class of 31 children recorded the maximum daily temperatures for the month of July with the following results. The median and quartiles are shown on the stem and leaf diagram. Draw the corresponding box plot and use it to identify any outliers.

Example



So, as a simple rule, take points lying more than 1.5 times the interquartile range above Q_3 or below Q_1 as 'outliers', and label them clearly as such:



Now the probability of a point lying to the right of this can be shown to be extremely small. Similarly, the probability of a point lying to the left of $\bar{x} - 2\frac{2}{3}s$ is also extremely small.

This point would be $2\frac{2}{3}s$ to the right of $\bar{x}$.

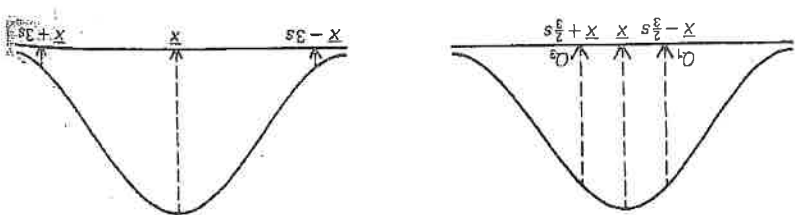
i.e. a point lying $1.5 \times \frac{5}{3}s = 2s$ to the right of Q_3 .

Now consider a point lying $1.5 \times (Q_3 - Q_1)$ to the right of Q_3 ,

$$= \frac{5}{3}s \quad \text{where } s \text{ is standard deviation}$$

$$= \bar{x} + \frac{5}{3}s - (\bar{x} - \frac{2}{3}s)$$

$$\text{The interquartile range} = Q_3 - Q_1$$



for a normal distribution the quartiles are situated at approximately $\frac{5}{3} \times$ standard deviation either side of the mean, and that nearly all values lie within 3 standard deviations of the mean.

Temperature in °F.

[illegible]
$$Q_1 = 8\text{th value} = 66^\circ\text{F}$$

$$Q_2 = 16\text{th value} = 70^\circ\text{F}$$

$Q_3 = 24^{\text{th}} \text{ value} = 73.5$

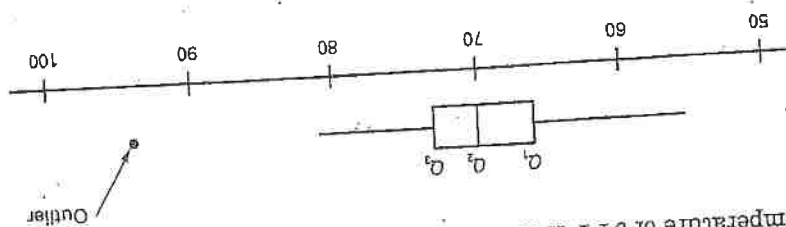
$$Q_3 - Q_1 = 73 - 66 = 7^\circ\text{F}$$

$$\text{Upper boundary} = 63 + 1.5 \times 7 = 73.5 + 10.5 = 83.5$$

$$\text{Lower boundary} = 61 - 1.5 \times 7 = 66 - 10.5 = 55.5$$

So it would appear that the temperature recorded as 94°F is an outlier. (It was probably recorded wrongly, since it is most unusual to have just one day with extremely high temperatures.) However, the temperature of 56°F is not an outlier.

The whiskers are drawn down to 56°F and up to 81°F and the temperature of 94°F is labelled as an outlier, as shown.



Handling data

Exercise

In Nos 1 - 10 criticise the given question and suggest a better alternative, if possible.

1 What is the size of your house?

2 What is your height?

3 Do you think that the town centre should be pedestrianised in order to reduce the number of accidents?

4 What is your marital status?
Ring the appropriate word.

Single Married

5 How many times have you been to the cinema during the past twelve months?
Tick the appropriate box.

0 - 5 5 - 12 12 - 24 24 - 52 Over 52

6 Who is your favourite pop group / artist?

7 What is your favourite football team?
Tick the appropriate box.

Cardiff City Swansea City Wrexham Newport AFC Other

8 Do you go to watch rugby matches?
Circle the appropriate answer.

Never Occasionally Often Regularly

9 Is your family income low, average or high compared with your neighbours?
Circle the appropriate answer.

Low Average High

10 How many cups of tea do you drink per day?

None 1 or 2 3 or 4 5 or 6 7 or 8 More than 8

55. Mean, Median, Mode and Range

You need to know what these 4 terms mean for your exam.

1. Mean – the mean is just the Average. The formula we use is:

$$\text{Mean} = \frac{\text{Total of items}}{\text{Number of items}}$$

Notice that this formula can be re-arranged as

$$\text{No. of Items} = \frac{\text{Total of items}}{\text{Mean}} \quad \text{or} \quad \text{Total of items} = \text{Mean} \times \text{Number of Items}$$

2. Median – the median is the middle value of a list of numbers that are arranged in order of size (ASCENDING ORDER).
Eg.) 1, 3, 5
The median here is 3

When there are 2 middle numbers the median is the average of these two middle numbers, i.e. the number halfway between the two middle numbers.

Eg.) 2, 6, 8, 10

$$\text{The median is } \frac{6+8}{2} = \frac{14}{2} = 7$$

Notice that

- when we had 3 numbers the median was the 2nd number.
- when we had 4 numbers the median was the 2.5th number.

In general if we have n numbers the median will be the $\frac{n+1}{2}$ number.

3. Mode – this is the value that occurs most often.
Eg.) 1, 1, 3, 3, 3, 3, 5, 10, 10
The mode is 3 because it occurs most often.

4. Range – This is the difference between the biggest and smallest value.
Eg.) -2, 3, 5, 5, 8, 9
The range is $9 - (-2) = 11$

Example
Find the mean, median, mode and range of these numbers:
2, 5, 3, 3, 2, 6, -4, 0, 9, -3, 1, 6, 3, -2, 3

Solution

We first arrange the numbers in ascending order:

-4, -3, -2, 0, 1, 2, 2, 3, 3, 3, 3, 5, 6, 6, 9

- Mean = $\frac{-4-3-2+0+1+2+2+3+3+3+3+5+6+6+9}{14} = 2.21$ (2 d.p.)

- There are 14 numbers so the median will be the $\frac{14+1}{2} = 7.5^{\text{th}}$ number. The 7th number is 2 and the 8th is 3, so the median will be the average of 2 and 3:

$$\frac{2+3}{2} = 2.5$$

- The mode is 3.
- Range = $9 - (-4) = 9 + 4 = 13$

Example
The mean weight of six people is 83kg. If three more people, weighing 93kg, 107kg and 78kg join them, what is their new mean weight?

Solution

The total weight of the 6 people is given by

$$\text{Total of items} = \text{Mean} \times \text{Number of Items}$$

$$= 83 \times 6$$

$$= 498 \text{ kg}$$

Hence the total weight of the 9 people is given by

$$\text{Total Weight} = 498 + 93 + 107 + 78 = 776 \text{ kg}$$

Thus the mean weight of the nine people is

$$\text{Mean} = \frac{776}{9} = 86.22 \text{ kg (2 d.p.)}$$

56. Frequency Tables

56.1. Data Types

Discrete Data is data that can be identified by a single number. Some examples are:

- the number of tys scored by a rugby team,
- the number of people with blue eyes,
- the number of sisters people have,
- the number of marks obtained in a test.

Continuous Data is data that cannot be measured exactly and is usually rounded off. Examples of continuous data include:

- the height of a number of people,
- the weight of a number of people,
- length of time
- a person's age

56.2. Frequency Tables

A frequency table is just a neat way of displaying data. Imagine you conducted a survey of the number of sisters people have.

After asking 46 people you found that:

- 7 has 0 sisters
- 15 had 1 sister
- 12 had 2 sisters
- 8 had 3 sisters
- 3 had 4 sisters
- 1 had 5 sisters
- 0 had 6 sisters

This may be better expressed in a table like the one below:

No. of Sisters	Frequency
0	7
1	15
2	12
3	8
4	3
5	1
6	0
Total	46

or

No. of sisters	Frequency
0	7
1	15
2	12
3	8
4	3
5	1
6	0
Total	46

The word FREQUENCY just means how many.

56.3. Mean, Mode, Median and Range

In order to calculate the mean from a frequency table we must add another column as shown below:

No. of sisters	Frequency	No. $\times$ Frequency
0	7	0
1	15	15
2	12	24
3	8	24
4	3	12
5	1	5
6	0	0
Total	46	80

To find the entries in the final column we just multiply the number of sisters by the frequency, for example in the 2nd row we have $2 \times 12 = 24$

Using this table we see that:

$$1. \text{ Mean} = \frac{\text{Total number of sisters}}{\text{Total number of people}} = \frac{46}{80} = 1.74 \text{ (Sisters per person)}$$

2. Median: There are 46 values in total so the median will be the $\frac{46+1}{2} = 23.5^{\text{th}}$ value, i.e. the number halfway between the 23rd and the 24th.

0000000 111111111111 222222222222 33333333 444 5

$\swarrow$ 23rd value $\searrow$ 24th value
 ▲

So the median is half way between 2 and 2, namely 2.

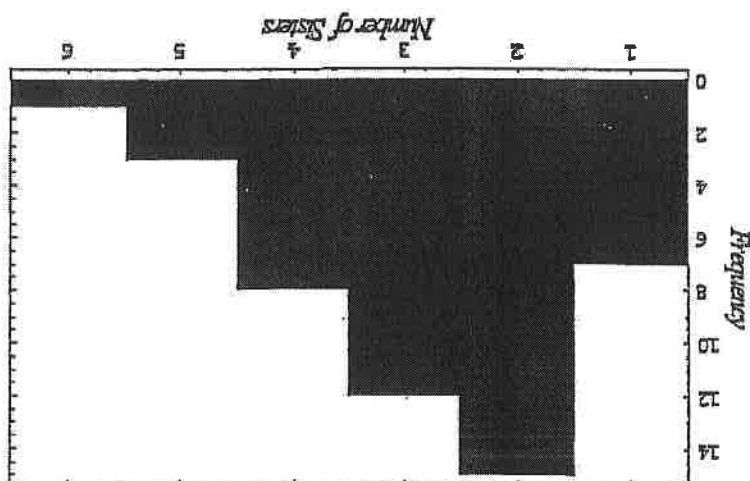
3. Mode: This is just the group with the most entries, namely 1.

4. Range: This is $5 - 0 = 5$
 Don't be tempted to say $6 - 0 = 6$. The 6 value does not occur!!!

56.4. Histograms

The data is a frequency table may be displayed graphically using a bar graph that we called a HISTOGRAM.

Drawing a histogram for the previous frequency table gives:

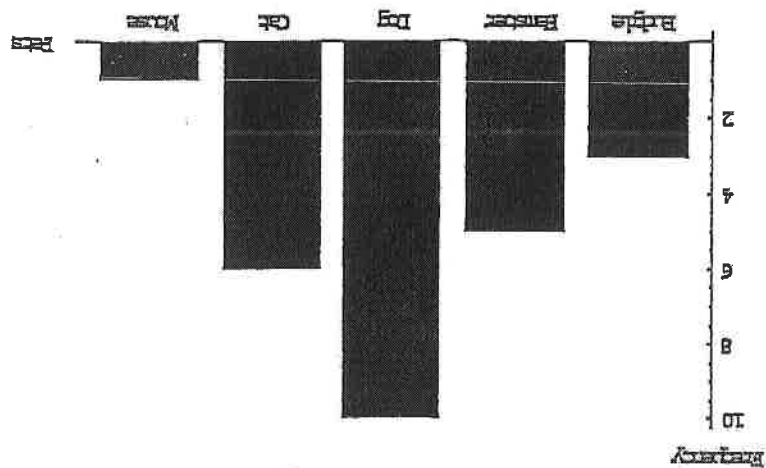


Example

A survey is conducted of the pets people have. The results were:

Pet	Budgie	Hamster	Dog	Cat	Mouse	Total
Frequency	3	5	10	6	1	25

Here we are dealing with totally separate items (a cat is completely different from a dog, etc.) so when you draw the histogram you leave gaps between the bars.



57. Frequency Distributions with Class Intervals

57.1. Grouped Frequency Tables

In a grouped frequency table, data is bunched together into groups called class intervals.

Example
The examination marks of 120 students are recorded and shown in the grouped frequency table below:

Mark	0-9	10-19	20-29	30-39	40-49	50-59	60-69
Freq	5	14	22	29	27	19	4

Modal Class or Modal Group
This is the class-interval that occurs most often, i.e. the one with the highest frequency. For the above example the modal class is 30-39 marks.

Mid Values

The centre of the marks 0-9 is the average of 0 and 9:

$$\frac{0+9}{2} = 4.5$$

The centre of the marks 10-19 is the average of 10 and 19:

$$\frac{10+19}{2} = 14.5$$

Etc.

Example
A survey of the ages of 40 people is conducted. The results are shown in the grouped frequency table below:

Age	0-9	10-19	20-29	30-39	40-49
Freq	6	15	10	6	3

The modal class here is 10-19.

Ages are usually given in completed years so somebody who is 9 yrs 11 months old would go in the 0-9 category. Thus the interval 0-9 is from 0 years to 10 years and the centre of the interval is

$$\frac{0+10}{2} = 5.$$

(Don't make the mistake of saying the centre is $\frac{0+9}{2} = 4.5$)

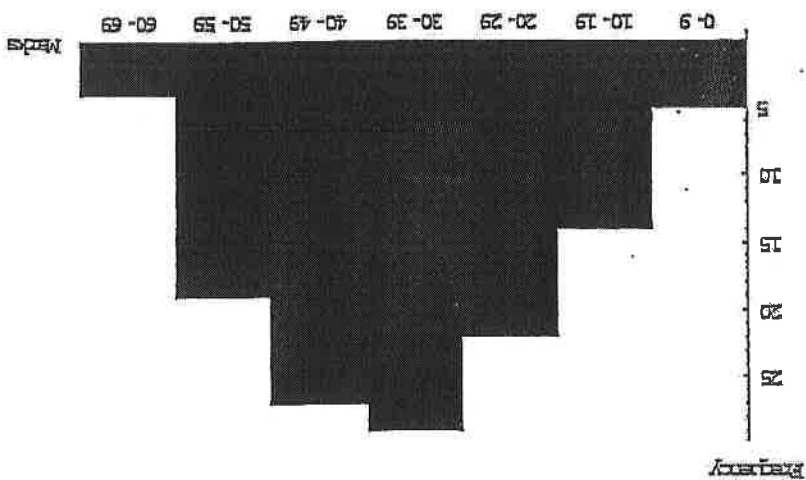
Doing this for each interval we see the midpoints are: 5, 15, 25, 35, 45.

57.2. Histograms

As with frequency tables we may express grouped frequency tables graphically by means of a (grouped) histogram.

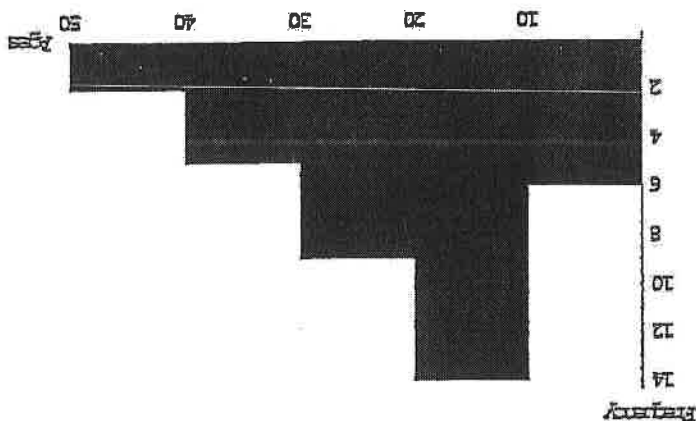
Example
Draw a histogram for the grouped frequency table showing the examination marks of 120 students.

Mark	0-9	10-19	20-29	30-39	40-49	50-59	60-69
Freq	5	14	22	29	27	19	4



Example
Draw a histogram for the grouped frequency table below showing the ages of 40 people.

Age	0-9	10-19	20-29	30-39	40-49
Freq	6	15	10	6	3



NB: You must be careful to start and finish each block at the correct point.

Example Consider the frequency table below, showing the heights of children measured to the nearest cm.

Height of Children (to nearest cm)	Frequency
120 - 129	2
130 - 139	5
140 - 149	23
150 - 159	55

Since these measurements are to the nearest cm the interval 120 - 129 is from 119.5 cm to 129.5 cm. This could be written as: $119.5 \leq h < 129.5$

Doing this for each interval the table becomes:

Height of Children (to nearest cm)	Frequency
$119.5 \leq h < 129.5$	2
$129.5 \leq h < 139.5$	5
$139.5 \leq h < 149.5$	23
$149.5 \leq h < 159.5$	55

Mid Points:

Mid point of 120 - 129 is

$$\frac{119.5 + 129.5}{2} = 124.5 \text{ cm}$$

Mid point of 130 - 139 is

$$\frac{129.5 + 139.5}{2} = 134.5 \text{ cm}$$

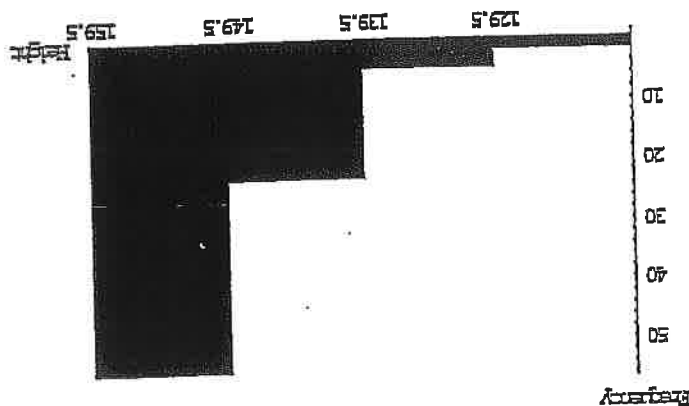
Mid point of 140 - 149 is

$$\frac{139.5 + 149.5}{2} = 144.5 \text{ cm}$$

Mid point of 150 - 159 is

$$\frac{149.5 + 159.5}{2} = 154.5 \text{ cm}$$

Drawing a histogram of the data gives:



NB: In your exam the height axis may be already drawn and marked with the values 120, 130, 140, 150. A common mistake is to draw the block for 120 - 129 from 120 to 130. This is WRONG!!! The block must be drawn from 119.5 to 129.5.

57.1. Estimating the Mean

When we have a frequency distribution with class intervals the best we can do is estimate the mean because we don't know the exact values. To do this we assume that each of the items in a class interval take the mid value of the class interval:

Example

Age	Frequency	Mid Value	Frequency × Mid Value
120-129	2	124.5	249
130-139	5	134.5	672.5
140-149	23	144.5	3323.5
150-159	55	155.5	8552.5
TOTALS	85		12797.5

- To find an estimate of the mean we use the formula:

$$\text{Mean} = \frac{\text{Total Age}}{\text{Number of People}} = \frac{12797.5}{85} = 150.56 \text{ (2 dp)}$$

- The modal group is the interval that occurs most often, namely 150 – 159
- The median can't be found. However we know that when the data is arranged in ascending order the median will be the $\frac{85+1}{2} = 43^{\text{rd}}$ value.

After 120 – 129 we have had 2 values.
 After 130 – 139 we have had 2 + 5 = 7 values.
 After 140 – 149 we have had 7 + 23 = 30 values.
 After 150 – 159 we have had 30 + 55 = 85 values.

Hence the 43rd value lies in the 150 – 159 groups. We say that the median lies in the 150-159 group.

58. Cumulative Frequency

Remember that frequency tells us how many times a particular result has occurred. We now consider cumulative frequency which tells us the number of times a result "less than" or "less than or equal to" a stated value has occurred. To create a cumulative frequency table from a frequency table we simply add successive frequencies to give a running total. This is best illustrated by means of an example:

Example 1

Draw a cumulative frequency table for the data below:

Wages (£w)	Frequency
$0 < w \leq 80$	7
$80 < w \leq 100$	15
$100 < w \leq 130$	41
$130 < w \leq 150$	20
$150 < w \leq 180$	7

So our cumulative frequency table would be

Wage	Cumulative Frequency
≤ 80	7
≤ 100	22
≤ 130	63
≤ 150	83
≤ 180	90

Example 2

Draw a cumulative frequency table for the data below:

Height (cm)	Frequency
141-150	4
151-160	9
161-170	20
171-180	33
181-190	36

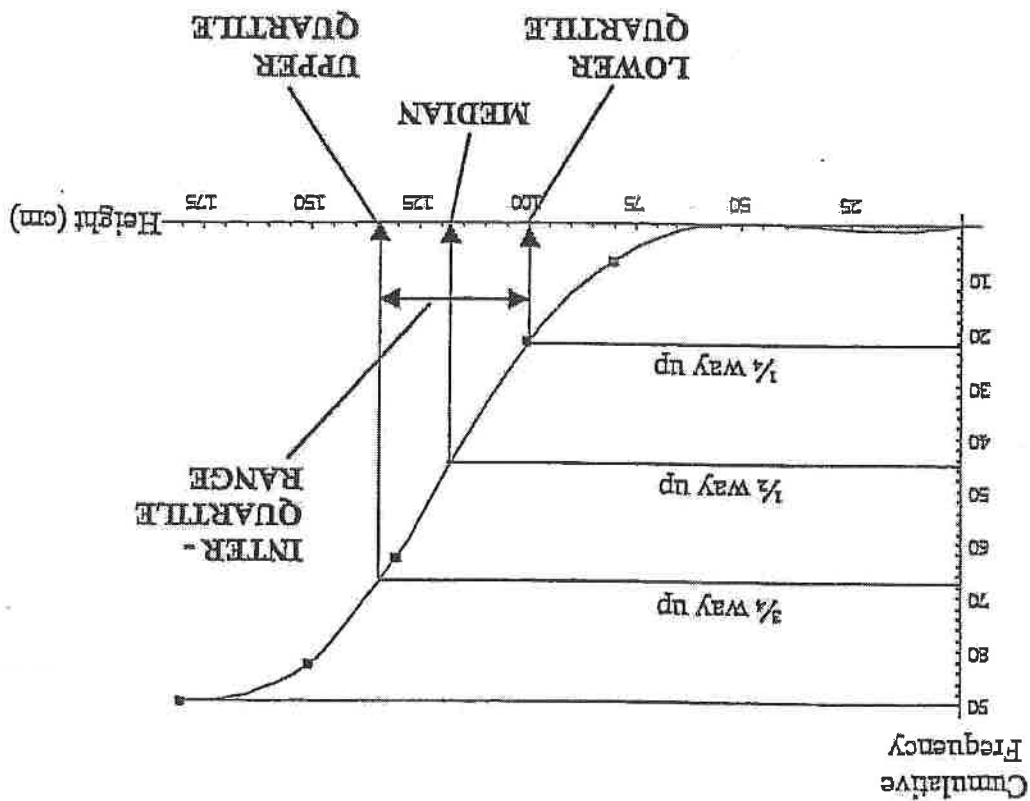
So our cumulative frequency table would be

Height (cm)	Cumulative Frequency
< 150.5	4
< 160.5	13
< 170.5	33
< 180.5	36
< 190.5	102

NB: Usually the values in the first row of your cumulative frequency table will be given, but if they are not then you should use the upper class interval of each class interval.

58.1. Cumulative Frequency Curve

You will also be required to plot this data on a graph and join the points with a smooth curve. The curve you obtain is called a Cumulative Frequency Curve. Plotting the curve for the table in Example 1 we obtain:



Note that from the cumulative frequency curve you can find the following quantities:

MEDIAN

Exactly halfway up, then across, then down and read off the bottom scale.

LOWER AND UPPER QUARTILES

Exactly $\frac{1}{4}$ and $\frac{3}{4}$ up the side, read across, then down and read off the bottom scale.

THE INTER-QUARTILE RANGE

This is the distance on the bottom scale between the lower and upper quartiles.

From the above cumulative frequency curve we see that:

$$\begin{aligned}\text{Upper Quartile} &= 135 \\ \text{Lower Quartile} &= 100\end{aligned}$$

$$\begin{aligned}\text{So Inter-Quartile Range} &= 135 - 100 \\ &= 35\end{aligned}$$

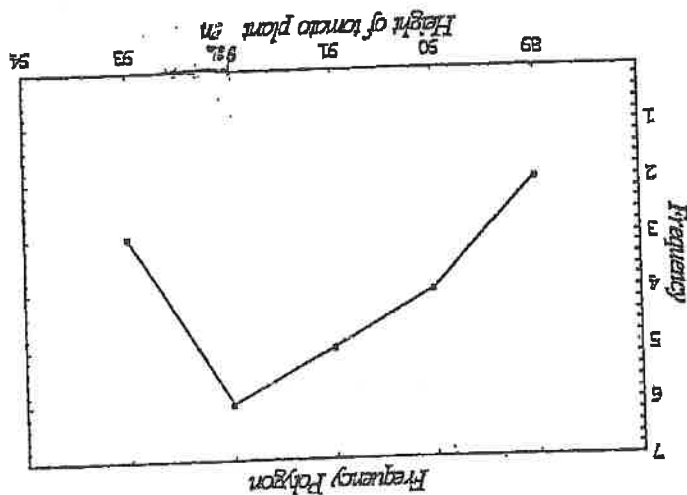
59. Frequency Polygons

A frequency polygon is a line graph that shows the shape of a frequency distribution. It is very easy to plot a frequency polygon for a frequency table as the following example illustrates.

Example Construct a frequency polygon for the following table showing the heights of 20 tomato plants:

Height of tomato plant	Frequency
89 cm	2
90 cm	4
91 cm	5
92 cm	6
93 cm	3

We plot "height of tomato plant" along the horizontal axis and "frequency" along the vertical axis. This gives:



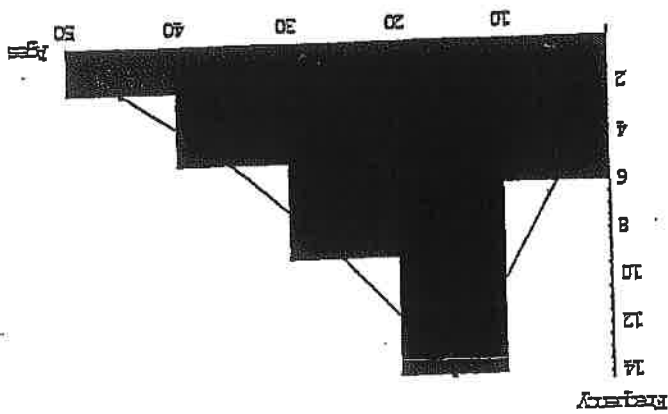
Frequency Always Goes on the Vertical Axis!!!!!!

59.1. Frequency Polygons for Grouped Frequency Data

Frequency polygons are more commonly used to show the shape of a grouped frequency distribution. If you already have a histogram for the data then a frequency polygon can be drawn by simply joining the mid-points of the tops of each bar.

Example

Age	Frequency
0-9	6
10-19	14
20-29	9
30-39	5
40-49	2
TOTAL:	36



If you don't already have a histogram then you must first work out the mid value of each group. You then plot these mid values against the frequency to obtain your frequency polygon.

Example
Construct a frequency polygon for the grouped frequency table showing the scores of 200 students in an examination:

Mark	0-10	11-20	21-30	31-40	41-50	51-60
Freq.	2	7	9	21	53	8

Solution

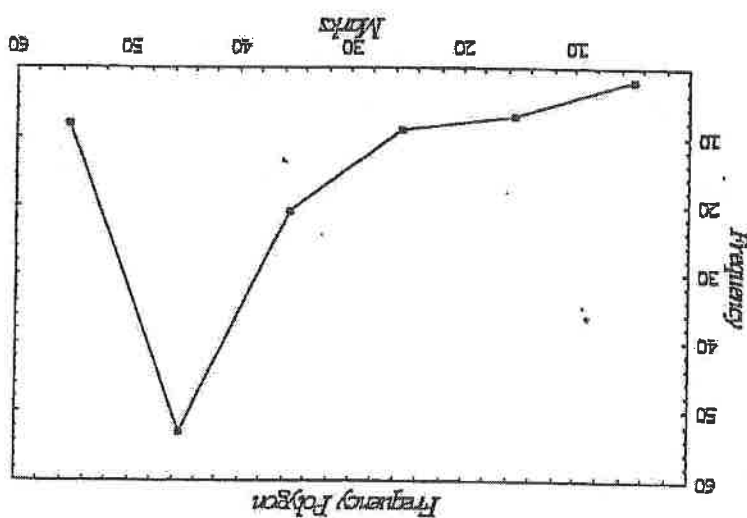
The mid value of 0-10 is 5

The mid value of 11-20 is $\frac{11+20}{2} = 15.5$

The mid value of 21-30 is $\frac{21+30}{2} = 25.5$

Etc.

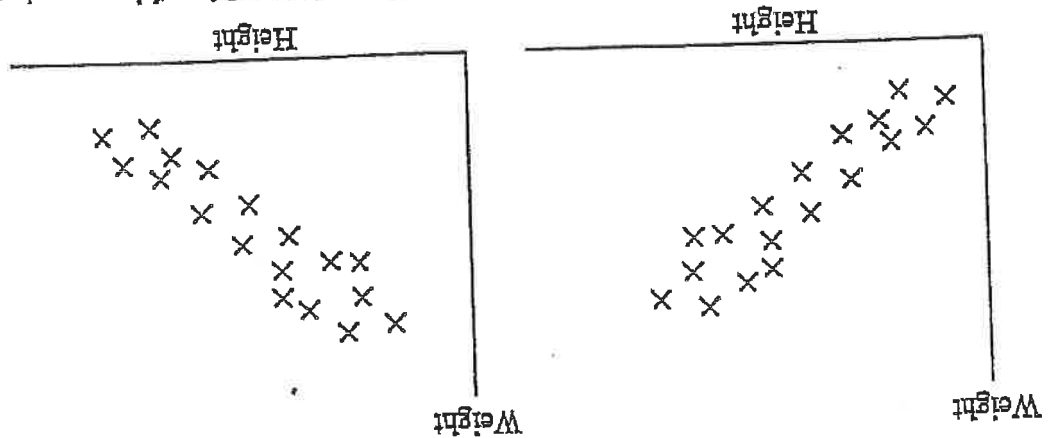
Doing this for each of the groups we see the mid values are
5, 15.5, 25.5, 35.5, 45.5, 55.5



Once again FREQUENCY goes on the vertical axis!!!!

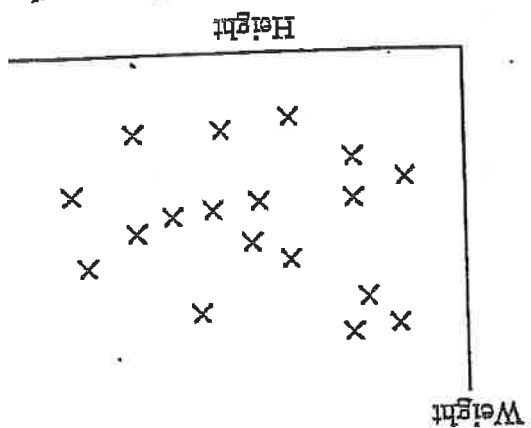
60. Scatter Diagrams

Imagine you wanted to investigate whether taller people weigh more than shorter people. You could collect the heights and weights of a group of people and compare them using a scatter diagram. You plot the height against the weight for each person and from the shape of your graph you can see if there is any relationship.



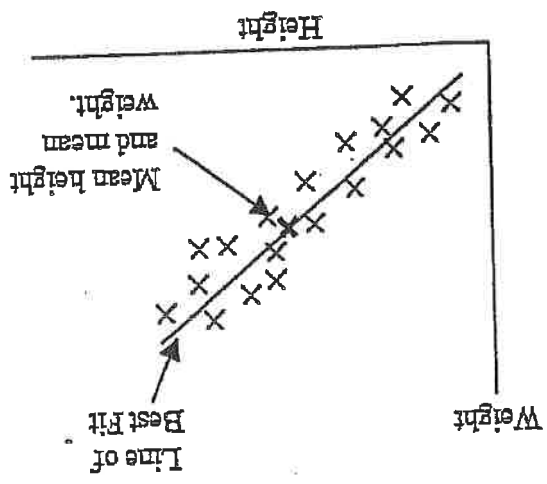
positive relationship: this suggests a taller person weighs more

negative relationship: this suggests a taller person weighs less



no relationship: we can't tell anything from the data.

If the points on the scatter diagram lie within a narrow band around an imaginary line then you can draw this line by eye.
This line is called the line of best fit. If you are told what the mean height and the mean weight is then your line of best fit must pass through this point.



The line of best fit is used to predict values for your data as we see in the following example.

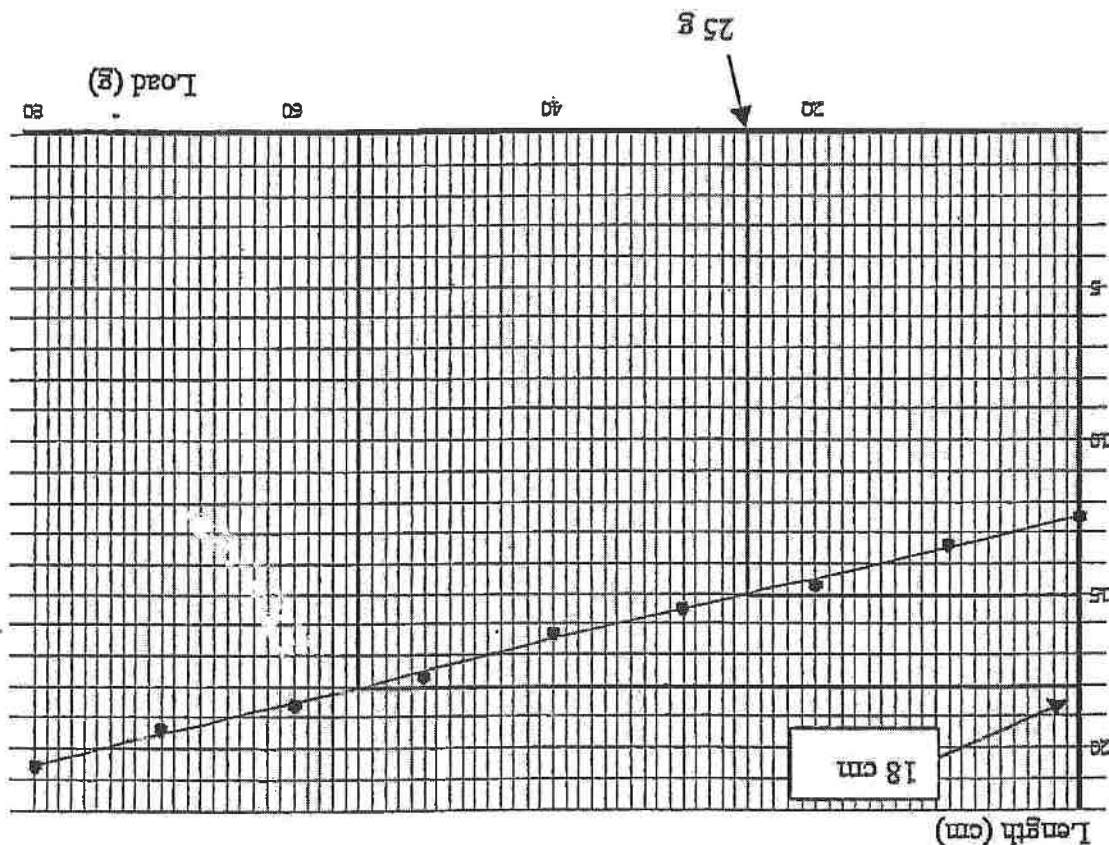
Example

A vertical spring was stretched by a weight at its lower end. The length of the spring for different loads was measured and the results recorded as follows:

Load (g)	0	10	20	30	40	50	60	70	80
Length (cm)	12.5	13.4	14.7	15.5	16.3	17.7	18.6	19.4	20.6

Draw a scatter diagram for this data and then draw a line of best fit. Use this to estimate:

1. The load when the length of the spring is 15cm
2. The length of the spring for a load of 55g.



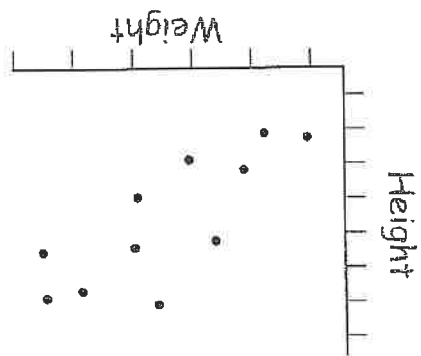
Using our line of best fit we see that:

1. When the length is 15cm the load is 25g.
2. When the load is 55g the length is 18cm.

Scatter Plots

A Scatter (XY) Plot has points that show the relationship between two sets of data.

In this example, each dot shows one person's weight versus their height.



(The data is plotted on the graph as "Cartesian (X,Y) Coordinates")

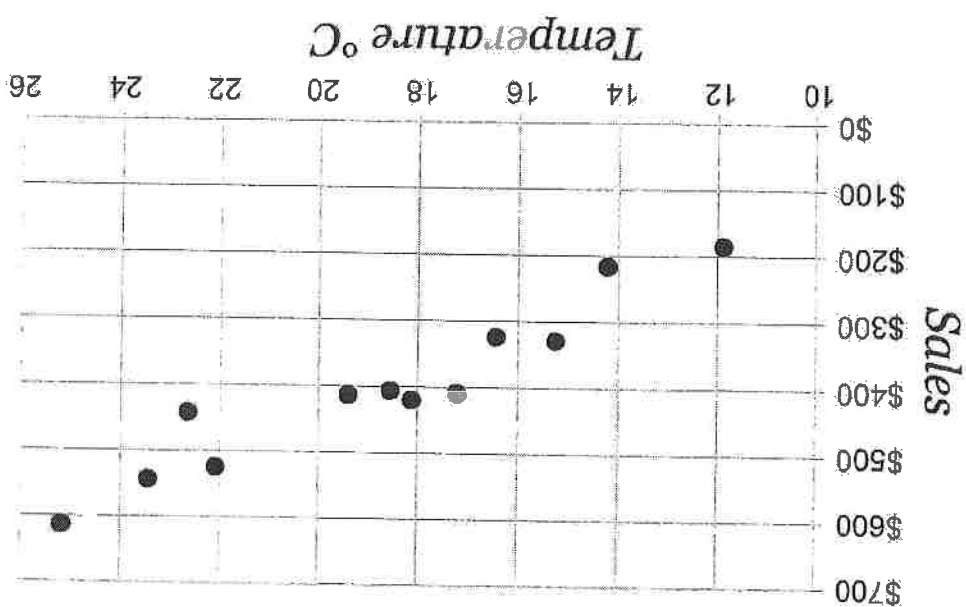
Example:

The local ice cream shop keeps track of how much ice cream they sell versus the noon temperature on that day. Here are their figures for the last 12 days:

Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	\$215
16.4°	\$325
11.9°	\$185
15.2°	\$332
18.5°	\$406
22.1°	\$522
19.4°	\$412
25.1°	\$614
23.4°	\$544
18.1°	\$421
22.6°	\$445
17.2°	\$408

69/

And here is the same data as a Scatter Plot:



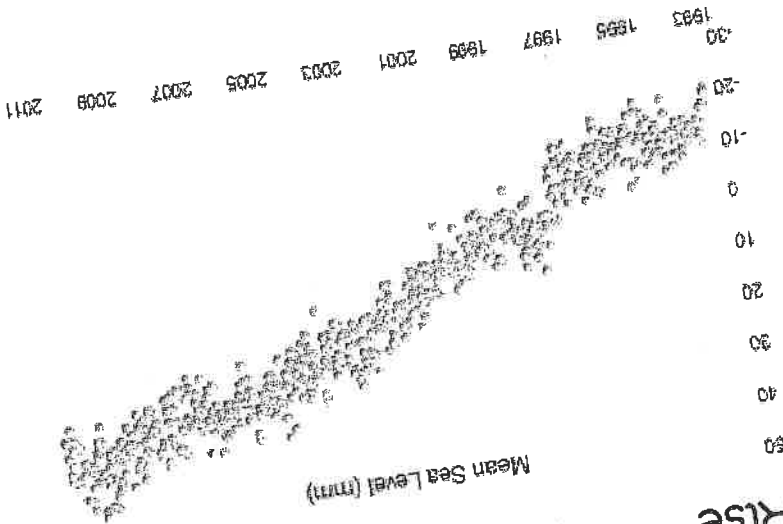
It is now easy to see that warmer weather leads to more sales, but the relationship is not perfect.

Line of Best Fit

We can also draw a "Line of Best Fit" (also called a "Trend Line") on our scatter plot:

71/

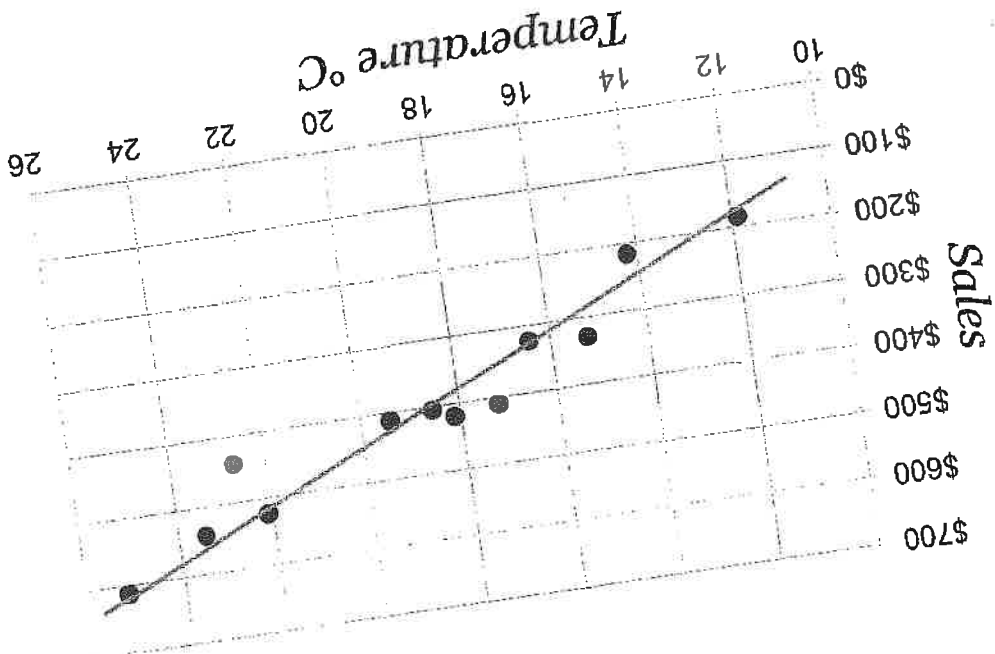
And here I have drawn on a
"Line of Best Fit".



A Scatter Plot of Sea Level
Rise:

Example: Sea Level Rise

Try to have the line as close as possible to all points, and as many points above the line as below.



$$\begin{aligned} \text{slope "m"} &= \frac{\text{change in y}}{\text{change in x}} \\ &= \frac{\$610 - \$180}{25^\circ - 12^\circ} \\ &= \frac{\$430}{13^\circ} \end{aligned}$$

First, find the slope:

(25°, \$610)

Let's estimate two points on the line near actual values: **(12°, \$180)** and

above

We can estimate a straight line equation from two points from the graph

Example: Straight Line Equation

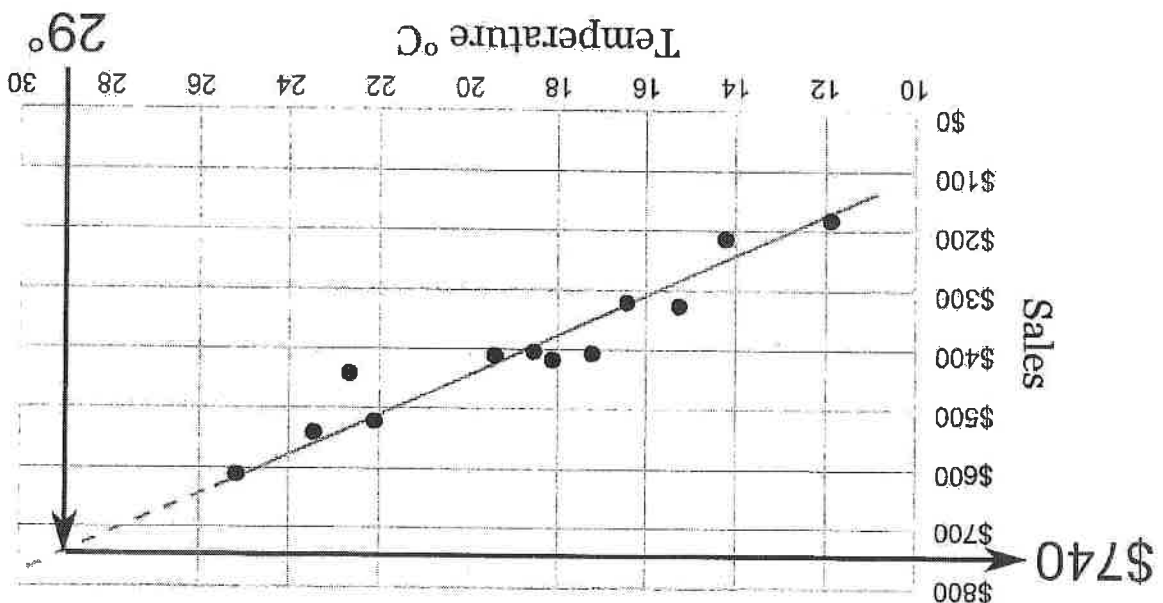
As well as using a graph (like above) we can create a formula to help us.

territory".

Careful: **Extrapolation** can give misleading results because we are in "uncharted

than any value we have).

Here we use **linear extrapolation** to estimate the sales at 29 °C (which is higher



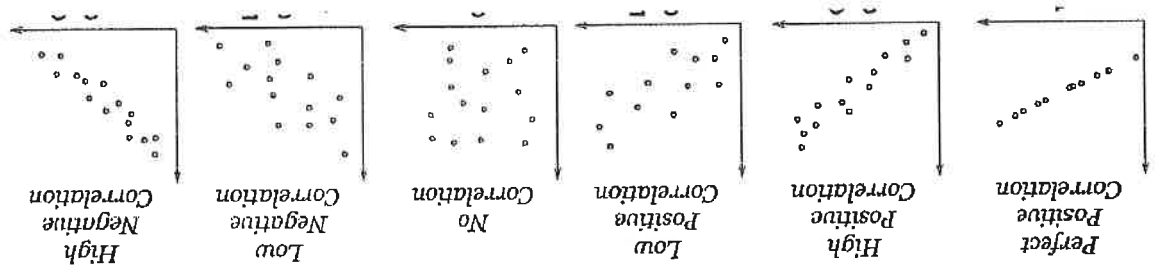
Correlation

When two sets of data are strongly linked together we say they have a **High Correlation**.

The word Correlation is made of **Co-** (meaning "together"), and **Relation**

- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases

Here we look at **linear correlations** (correlations that follow a line).



Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

The value shows **how good the correlation is** (not how steep the line is), and if it is positive or negative.

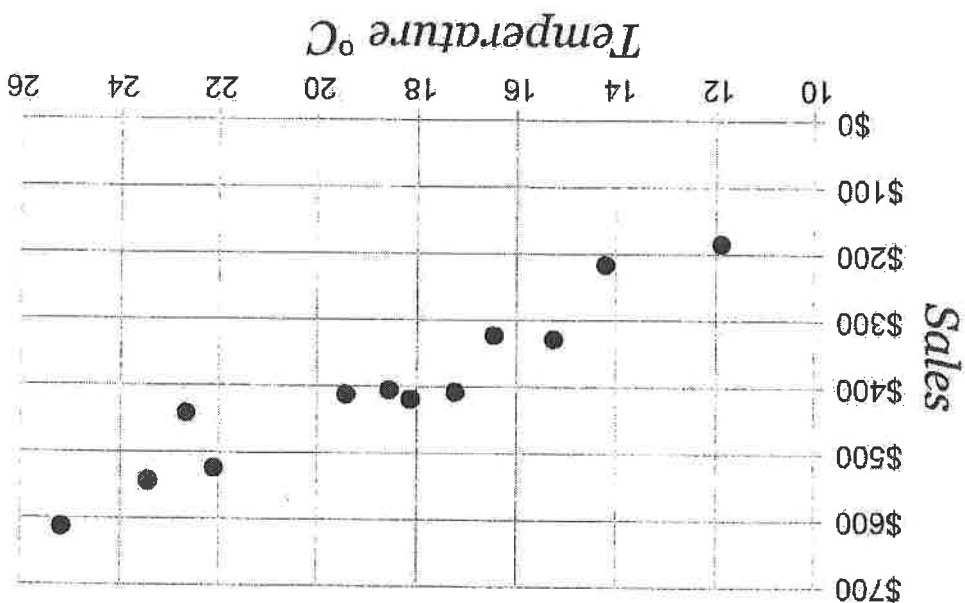
Example: Ice Cream Sales

The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days:

Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales

14.2°	\$215
16.4°	\$325
11.9°	\$185
15.2°	\$332
18.5°	\$406
22.1°	\$522
19.4°	\$412
25.1°	\$614
23.4°	\$544
18.1°	\$421
22.6°	\$445
17.2°	\$408

And here is the same data as a Scatter Plot :



We can easily see that warmer weather and higher sales go together. The relationship is good but not perfect.

In fact the correlation is **0.9575** ... see at the end how I calculated it.

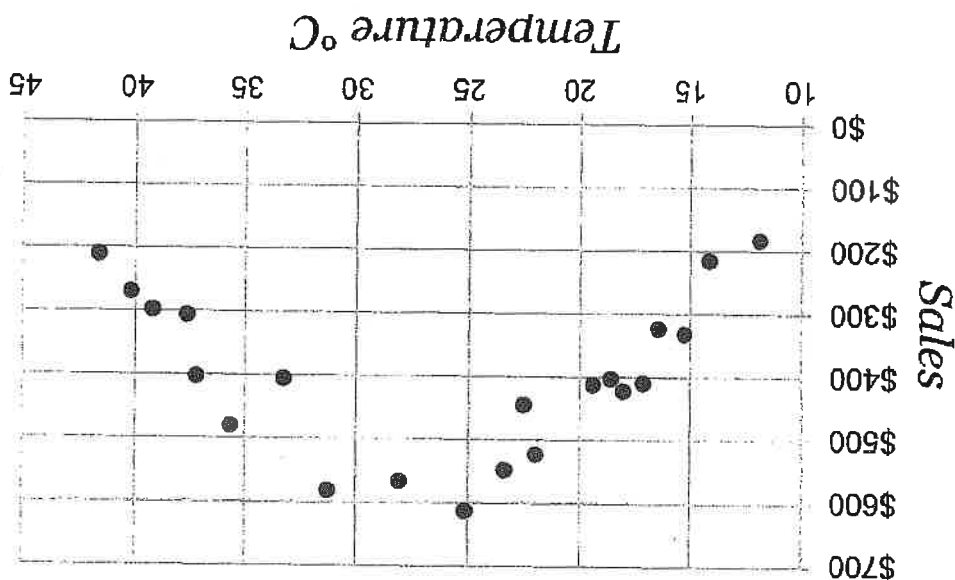
Correlation Is Not Good at Curves

The correlation calculation only works well for relationships that follow a straight line.

Our Ice Cream Example: **there has been a heat wave!**

It gets so hot that people aren't going near the shop, and **sales start dropping.**

Here is the latest graph:



The correlation value is now **0**: "No Correlation" ... !

The **calculated** correlation value is **0** (I worked it out), which means "no correlation".

But we can see the data **does have a correlation**: it follows a nice **curve** that reaches a peak around 25° C.

But the linear correlation calculation is not "smart" enough to see this.

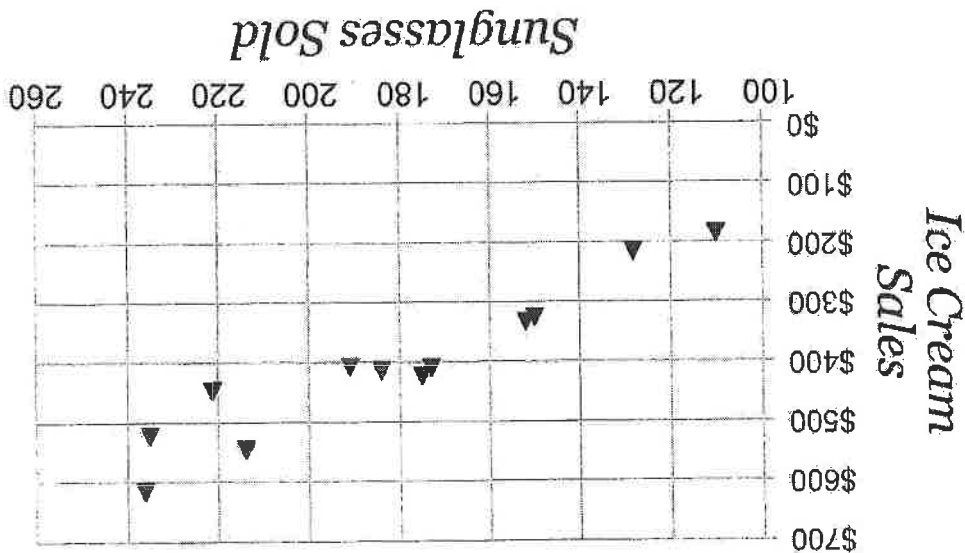
Moral of the story: make a Scatter Plot, and look at it! You may see a correlation that the calculation does not.

Correlation Is Not Causation

"Correlation Is Not Causation" ... which says that a correlation does **not** mean that one thing causes the other (there could be other reasons the data has a good correlation).

Example: Sunglasses vs Ice Cream

Our Ice Cream shop finds how many sunglasses were sold by a big store for each day and compares them to their ice cream sales:



The correlation between Sunglasses and Ice Cream sales is high

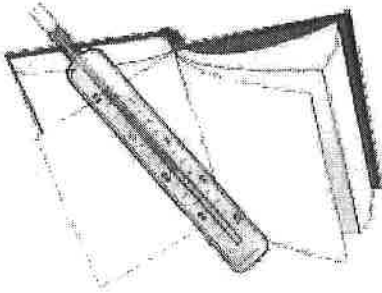
Does this mean that sunglasses make people want ice cream?

Example: A Real Case!

A few years ago a survey of employees found a strong positive correlation between "Studying an external course" and Sick Days.

Does this mean:

- Studying makes them sick?
- Sick people study a lot?



- Or did they lie about being sick to study more?

Without further research we can't be sure why.

How To Calculate

How did I calculate the value **0.9575** at the top?

I used "Pearson's Correlation". There is software that can calculate it, such as the CORREL() function in Excel or LibreOffice Calc ...

... but here is how to calculate it yourself:

Let us call the two sets of data "X" and "Y" (in our case Temperature is X and Ice Cream Sales is Y):

- Step 1: Find the mean of X, and the mean of Y
- Step 2: Subtract the mean of X from every X value (call them "a"), do the same for Y (call them "b")
- Step 3: Calculate: $a \times b$, a^2 and b^2 for every value
- Step 4: Sum up $a \times b$, sum up a^2 and sum up b^2
- Step 5: Divide the sum of $a \times b$ by the square root of [(sum of a^2) $\times$ (sum of b^2)]

Here is how I calculated the first Ice Cream example (values rounded to 1 or 0 decimal places):

- Σ is Sigma, the symbol for "sum up"
- $(x - \bar{x})$ is each x-value minus the mean of x (called "a" above)
- $(y - \bar{y})$ is each y-value minus the mean of y (called "b" above)

As a formula it is:

1 Calculate Means

Temp °C	Sales
18.7	\$402
17.2	\$408
22.6	\$445
18.1	\$421
23.4	\$544
25.1	\$614
19.4	\$412
22.1	\$522
18.5	\$406
15.2	\$332
11.9	\$185
16.4	\$325
14.2	\$215

2 Subtract Mean

"a"	"b"
-1.5	\$6
3.9	\$43
-0.6	\$19
4.7	\$142
6.4	\$212
0.7	\$10
3.4	\$120
-0.2	\$4
-3.5	-\$70
-6.8	-\$217
-2.3	-\$77
-4.5	-\$187

3 Calculate ab , a^2 and b^2

a^2	b^2	ab
2.3	36	-9
15.2	1,849	168
0.4	361	-11
22.1	20,164	667
41.0	44,944	1,357
0.5	100	7
11.6	14,400	408
0.0	16	-1
12.3	4,900	245
46.2	47,089	1,476
5.3	5,929	177
20.3	34,969	842

4 Sum Up

Σab	Σa^2	Σb^2
5,325	177.0	174,767

Note for Programmers

You can calculate it in one pass through the data. Just sum up x , y , x^2 , y^2 and xy (no need for a or b calculations above) then use the formula:

$$r_{xy} = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

Other Methods

There are other ways to calculate a correlation coefficient, such as "Spearman's rank correlation coefficient", but I prefer using a spreadsheet like above.

61. Probability

61.1. Introduction

Imagine an experiment is performed, e.g. a coin is tossed or a die rolled. Mathematicians call the list of all possible outcomes of the experiment a sample space.

Example

If we rolled a die the possible outcomes would be
 $S = \{1, 2, 3, 4, 5, 6\}$.

In probability we will be looking at the likelihood of an event occurring when an experiment is performed. Some examples of experiments and events you could perform are shown below:

Experiment	Event
Tossing a coin	Getting a head
Tossing a coin	Getting a tail
Rolling a die	Getting an even number
Rolling a die	Getting a 1

Example

Consider the experiment "tossing a coin" and the event "getting a head".

Possible outcomes on tossing a coin are $S = \{H, T\}$
 Ways you can get a head are $E = \{H\}$

Example

Consider the experiment "rolling a die" and the event "getting an odd number".

Possible outcomes on rolling a die are $S = \{1, 2, 3, 4, 5, 6\}$
 Ways you can get an odd number are $E = \{1, 3, 5\}$

To work out the probability of an event E occurring we use the formula:

$$\text{Probability of event } E = \frac{\text{Number of items in } E}{\text{Number of items in } S}$$

A short way to write "Probability of event E " is $P(E)$.

Example

What is the probability of getting an odd number when a die is rolled?

Solution

Possible outcomes on rolling a die are
 Ways you can get an odd number are

$S = \{1, 2, 3, 4, 5, 6\}$
 $E = \{1, 3, 5\}$

$$\text{So } P(E) = \frac{3}{6} = \frac{1}{2}$$

Example

What is the probability of picking an even number from 1 to 20 inclusive?

Solution

Possible outcomes are

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$$

Ways you can get an odd number are

$$E = \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$$

Hence

$$P(\text{even number}) = \frac{10}{20} = \frac{1}{2}$$

61.2. Certain Events

Probabilities always lie between 0 and 1. The bigger the probability the more likely the event is to occur.

If an event E is certain to occur then $P(E) = 1$
If an event E definitely won't occur then $P(E) = 0$

Example

A coin is tossed what is the probability that

a) a head or tail will occur?

b) a 3 will occur

Solution

a) $P(\text{head or tail}) = 1$

b) $P(3 \text{ occurring}) = 0$

61.3. Combining Events

Let A and B be two events. You need to know the following two formulae:

• $P(A \text{ OR } B) = P(A) + P(B)$

OR $\rightarrow$ ADD

• If A and B are independent (this means the result of one doesn't affect the other)

$$P(A \text{ AND } B) = P(A) \times P(B)$$

AND $\rightarrow$ MULTIPLY

Example

A die is rolled twice. What is the probability of obtaining an even number on your first roll followed by a 3 on your second roll?

Solution

Let A = even number

$$\text{So } P(A) = \frac{6}{2} = \frac{3}{1}$$

Let B = obtain 3

$$\text{So } P(B) = \frac{1}{6}$$

Hence

$$P(A \text{ and } B) = P(A) \times P(B) = \frac{3}{1} \times \frac{1}{6} = \frac{1}{2}$$

61.4. Complementary Events

Imagine that we know the probability of an event E occurring. To find the probability of E not occurring we use the formula:

$$P(E \text{ not occurring}) = 1 - P(E)$$

Example

The probability of obtaining a 3 when a die is rolled is $\frac{1}{6}$. What is the probability of

not obtaining a 3?

Solution

$$P(\text{not obtaining } 3) = 1 - \frac{1}{6} = \frac{5}{6}$$

61.5. Combined Events

Two or more events may occur together (e.g. we could toss a coin and roll a die at the same time). This gives a combined or compound event. A good way to show all possible outcomes when two events occur is to use a table.

Example

David tosses a coin and Bob throws a die.

- What is the probability of a head on the coin and 5 on the die?
- What is the probability of getting a tail and an even number?

Solution

First we tabulate all possible outcomes:

Coin		Die					
H	T	1	2	3	4	5	6
H1	T1	H2	T2	H3	T3	H4	T4
H5	T5	H6	T6				

The total number of outcomes is 12.

- The event E we are interested in is the outcome "a head and a 5". This can only happen in one way and has been shaded in yellow. So

$$P(\text{a head and a } 5) = \frac{1}{12}$$

- The event E we are interested in is a tail and an even number. These are circled in blue and we see that it can occur in 3 ways. So

$$P(\text{tail and an even number}) = \frac{3}{12} = \frac{1}{4}$$

Solution a)

Example
Three black cards are numbered 1 to 3 and four red cards are numbered 1 to 4. A black card is chosen at random followed by a red card.
a) List all the possible outcomes
b) What is the probability of getting two odd numbers?
c) What is the probability of the two cards adding up to 6?

<u>Black Card</u>		<u>Red Card</u>	
1	2	3	4
1	(1, 2)	(1, 3)	(1, 4)
2	(2, 1)	(2, 2)	(2, 3)
3	(3, 1)	(3, 2)	(3, 3)
			(3, 4)

b) We have shaded all possibilities of getting two odd numbers and see that there are 4. Thus

$$P(\text{two odd numbers}) = \frac{4}{12} = \frac{1}{3}$$

c) Here we draw a similar table to that in a) but this time we work out the sum:

<u>Black Card</u>		<u>Red Card</u>	
1	2	3	4
1	2	3	4
2	3	4	5
3	4	5	6
			7

The cards can add up to 6 in 2 ways. Hence we have

$$P(\text{two cards adding up to 6}) = \frac{2}{12} = \frac{1}{6}$$

62. Tree Diagrams

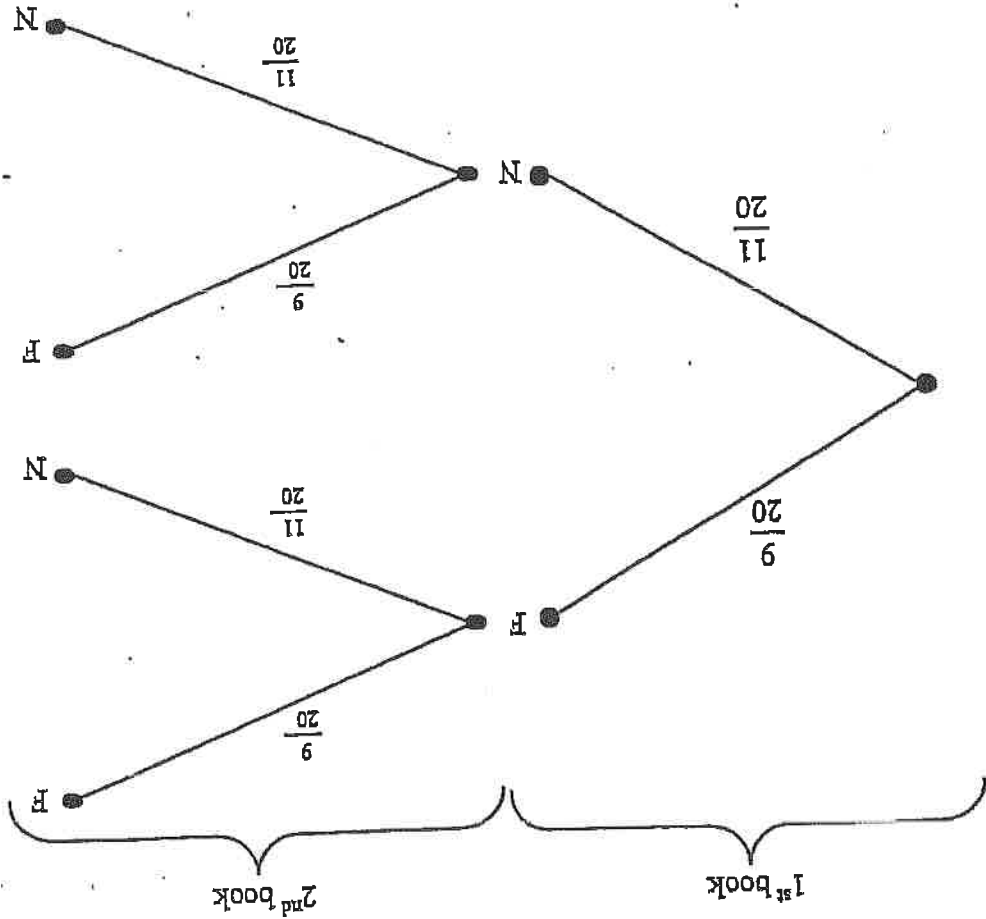
A probability tree diagram can be used to solve problems involving combined events. We illustrate by means of 2 examples:

Example 1

There are 20 books on a shelf in a room. 9 are fiction and 11 non-fiction. A person takes a book at random and looks at it. They then replace the book on the shelf and choose a book again at random. Draw a probability tree diagram to show the possible outcomes. What is the probability that the books taken are;

- both fiction,
- both non-fiction,
- one of each kind.

Solution



- There is only one route that gives us fiction AND fiction so we have

$$P(\text{fiction and non-fiction}) = \frac{9}{20} \times \frac{9}{19} = \frac{81}{380}$$
- There is only one route that gives us non-fiction AND non-fiction so we have

$$P(\text{non-fiction and non-fiction}) = \frac{11}{20} \times \frac{10}{19} = \frac{110}{380}$$

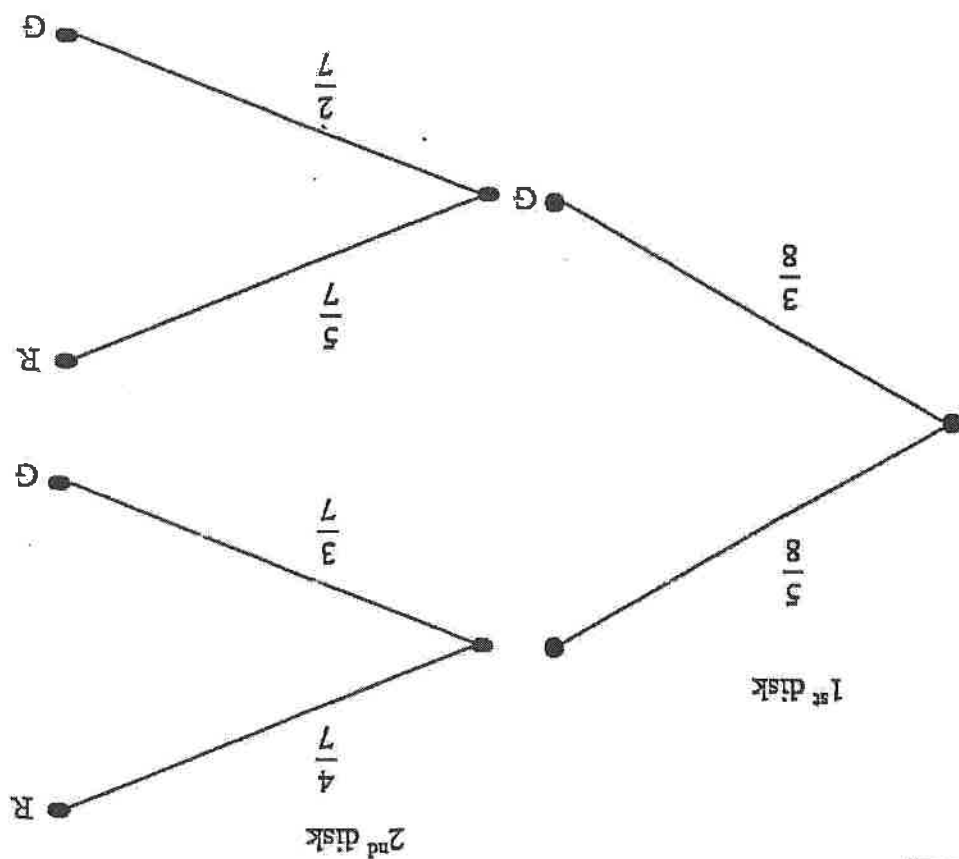
c) We could have fiction AND non-fiction OR non-fiction AND fiction. We work out the probabilities of each of these routes and add them:

$$P(\text{one of each kind}) = \frac{9}{20} \times \frac{11}{20} + \frac{11}{20} \times \frac{9}{20} = \frac{198}{400}$$

Example 2

A box contains 5 red disks and 3 green disks. Two disks are taken **WITHOUT REPLACEMENT**. Draw a tree diagram and hence find the probability that both disks are the same colour.

Solution



$$P(2 \text{ Reds}) = \frac{5}{8} \times \frac{4}{7} = \frac{20}{56} = \frac{5}{14}$$

$$P(2 \text{ Greens}) = \frac{3}{8} \times \frac{2}{7} = \frac{6}{56} = \frac{3}{28}$$

$$P(\text{Same colour}) = P(2 \text{ Reds or } 2 \text{ Greens})$$

$$= P(2 \text{ Reds}) + P(2 \text{ Greens})$$

$$= \frac{5}{14} + \frac{3}{28}$$

$$= \frac{10}{28} + \frac{3}{28}$$

$$= \frac{13}{28}$$

63. Relative Frequency

Suppose we want to find the probability that a drawing pin will land with the pin pointing up. One way to do this is to collect experimental data by throwing the drawing pin a number of times and recording the relative frequency, i.e. the number of times the pin lands pointing up divided by the total number of throws. The results might look like:

Number of times lands pointing up	Number of throws	Relative frequency
43	100	$\frac{43}{100} = 0.43$
64	200	$\frac{64}{200} = 0.32$
114	300	$\frac{114}{300} = 0.38$
136	400	$\frac{136}{400} = 0.34$
160	500	$\frac{160}{500} = 0.32$
170	600	$\frac{170}{600} = 0.283...$
203	700	$\frac{203}{700} = 0.29$
252	800	$\frac{252}{800} = 0.315$
274	900	$\frac{274}{900} = 0.304...$
310	1000	$\frac{310}{1000} = 0.31$
336	1100	$\frac{336}{1100} = 0.305...$
355	1200	$\frac{355}{1200} = 0.295...$

From the table it appears that the relative frequency is approaching the value 0.3. This is defined as the experimental probability.

We say

$$P(\text{pin lands facing up}) \approx 0.3$$

The sign $\approx$ means "approximately equal to".

For a large number of trials

$$P(A) \approx \frac{\text{Number of times A occurs}}{\text{Number of trials}}$$

This formula can be rearranged so that

$$\text{Number of times A occurs} \approx \text{Number of trials} \times P(A)$$

Example

Use relative frequency to estimate the following:

- How many tails would you expect to get in 20 tosses of a coin?
- How many times would you expect to get a 6 if you rolled a fair die 18 times?
- How many Hearts would you expect if you were dealt 12 cards at random?

Solution

- The probability is $\frac{1}{2}$. So we would expect to get $\frac{1}{2} \times 20 = 10$ tails.
- The probability of a 6 is $\frac{1}{6}$. So we would expect to get $\frac{1}{6} \times 18 = 3$ sixes
- The probability of a Heart is $\frac{1}{4}$. So we would expect to get $\frac{1}{4} \times 12 = 3$ Hearts

appearing.

appearing.

Example
Raffle tickets are sold in aid of the local church. 500 are blue, 200 white and 50 pink. They are all put into a box and taken out at random.

- That is the probability that the first ticket is blue?
- If there are 12 prizes to be won, about how many prize-winners would you expect to have a blue ticket?

Solution

$$a) P(B) = \frac{500}{750} = \frac{2}{3}$$

- Number of times blue occurs $\approx \text{Number of trials} \times P(B) = 12 \times \frac{2}{3} = 8$
We would expect 8 prize winners to have a blue ticket.

PROBABILITY EXERCISE 1

1. There are 16 girls in a class of 24 students. If one of the students is chosen at random what is the probability that the chosen student is:
 - a) male b) female.
2. The letters which make up the word TELEVISION are written on separate pieces of card and placed in a bag.
 - One of the cards is then selected at random. What is the probability that the chosen card is :
 - a) a vowel b) a consonant.
3. Peter has a bag containing 30 pieces of lego. 8 are red, 10 yellow and the rest are black. Jim draws out a piece of lego what is probability that the chosen piece is: a) red b) black c) white.
4. Rita chooses a month at random from the twelve months in the year. What probability is there that the chosen month
 - a) begins with the letter j b) begins with a vowel?
- 5) A bag contains 10 balls, 3 are red the rest are white.
 - If a ball is drawn out at random without replacement what is the probability that the chosen ball is not red?
- 6) A fair tetrahedral die is rolled once what chance is there of obtaining a 3?
- 7) A card is drawn from an ordinary deck of playing cards. what is the probability of obtaining
 - a) a queen b) a four c) the jack of diamonds.
- 8) The probability of the sun shining in Wales tomorrow is 0.1. calculate the probability of it raining tomorrow.
- 9) In a family board game a fair cubical die is rolled 180 time, how many times would you have expected a 6 to have turned up?
- 10) The probability that Neath beat Llanelli in rugby is 0.4, if they play 100 game how many games would you expect Llanelli to win?
- 11) Two fair coins are thrown in simultaneously, what is the chance of obtaining 2 heads?
- 12) A fair cubical die is rolled twice in succession, what chance is there of obtaining two 2's?

PROBABILITY EXERCISE 2

1. A fair coin is tossed 3 times in succession find, the probability of obtaining:
a) 3 tails b) a tail followed by a head followed by a head.
2. Two fair tetrahedral dice are rolled simultaneously. Calculate the probability of obtaining two 4's.
3. A bag contains 10 balls. 6 are red and 4 are white. A ball is drawn out at random its colour noted but not replaced. Then a second ball is drawn out. Calculate the probability of obtaining:
a) two red balls b) two balls of different colours will be chosen.
4. A bag contains 10 balls, 7 are red 3 are white. A ball is chosen at random its colour noted and then replaced, then a second ball is drawn out. calculate the probability of obtaining:
a) two white balls b) a ball of either colour.
5. A man tosses a fair coin, if he obtains a head he picks a ball from bag A which contains 3 red and 7 white balls. If he obtains a tail he picks a ball from bag B which contains 7 red and 3 white balls.
calculate the probability of obtaining a white ball.
6. A card is drawn at random from an ordinary pack of playing cards, not replaced then a second card is chosen.
Calculate the probability of obtaining ACE and a KING.
7. Two letters are selected at random from the word MISSISSIPPI. Find the probability that they are the same.
8. A bag contains 3 black, 3 white and 3 red balls. A random sample of two balls is drawn at random without replacement. Calculate the probability that the balls have different colours.

BASIC SET THEORY

A set is a group of objects. Each object is known as a member of the set. A set can be represented using curly brackets. So a set containing the numbers 2, 4, 6, 8, 10, ... is: $\{2, 4, 6, 8, 10, \dots\}$. Sets are often also represented by letters, so this set might be $E = \{2, 4, 6, 8, 10, \dots\}$. Alternatively, $E = \{\text{even numbers}\}$.

Common Sets

Some sets are commonly used and so have special notation:

 The set of integers

 The set of natural numbers (positive integers)

 The set of rational numbers

 The set of real numbers

For example:
 $x \in \mathbb{Z}$

means that x is in the set of integers,
i.e. x is an integer.

$\in$ means 'is a member of'

Other Notation

Subsets

If A is a subset of B , then all of the elements of A are also in B . For example, if $A = \{1, 2, 3\}$ and $B = \{1, 2, 3, 4, 5\}$ then $A \subseteq B$ ($\subseteq$ means is a subset of).

Number of Members

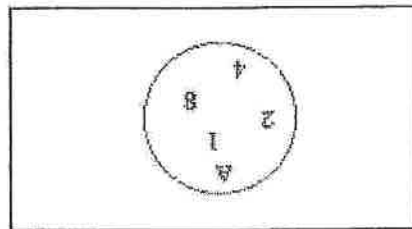
If $A = \{1, 2, 4, 8\}$, then $n(A) = 4$. This is because $n(A)$ means the number of members in set A .

The Universal Set

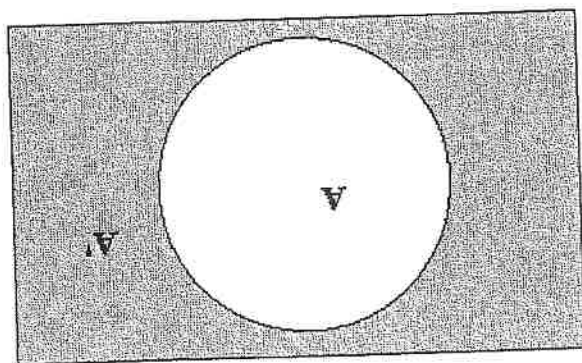
The universal set is the set of all sets. All sets are therefore subsets of the universal set.

Venn Diagrams

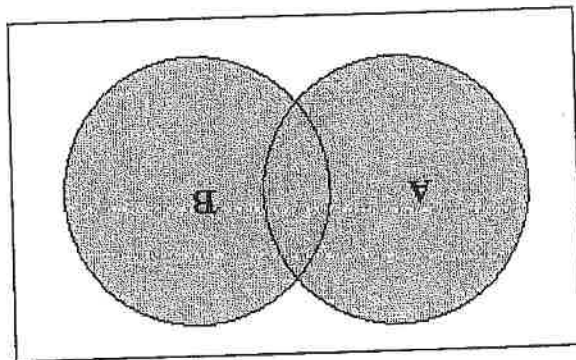
Venn diagrams are used to represent sets. Here, the set $A = \{1, 2, 4, 8\}$ is shown using a circle. In Venn diagrams, sets are usually represented using circles. The universal set is the rectangle. The set A is a subset of the universal set and so it is within the rectangle.



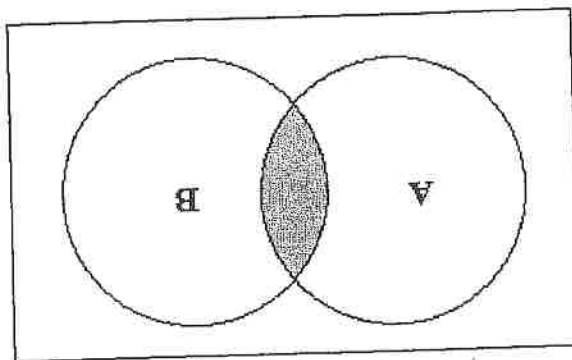
The complement of A , written A' , contains all events in the sample space which are not members of A . A and A' together cover every possible eventuality.



$A \cup B$ means the **union** of sets A and B and contains all of the elements of both A and B. This can be represented on a Venn Diagram as follows:



$A \cap B$ means the **intersection** of sets A and B. This contains all of the elements which are in both A and B. $A \cap B$ is shown on the Venn Diagram below:

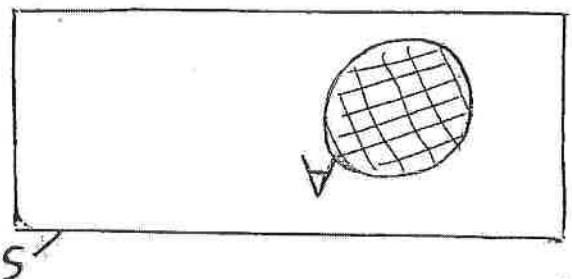


An important result connecting the number of members in sets and their unions and intersections is:

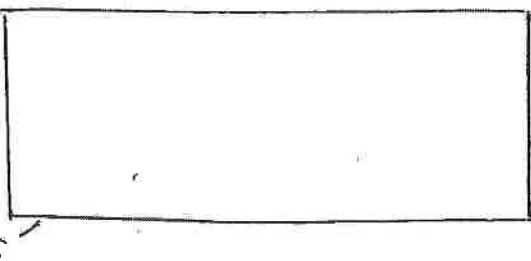
- $n(A) + n(B) - n(A \cap B) = n(A \cup B)$

VENN DIAGRAMS

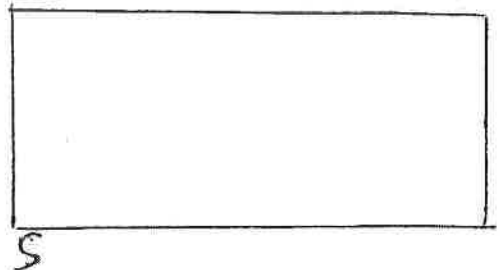
WE CAN DEMONSTRATE SAMPLE SPACES, EVENTS, INTERSECTION AND UNIONS USING VENN DIAGRAMS.
 THE SAMPLE SPACE (S) IS DEPICTED BY A RECTANGLE
 AND EVENTS ARE REPRESENTED BY CIRCLES
 WE CAN FIND PROBABILITIES BY SHADING THE REQUIRED AREA.



$P(A)$

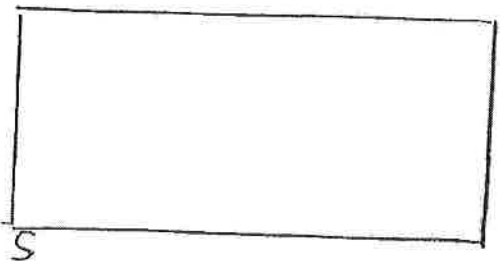


$P(A')$

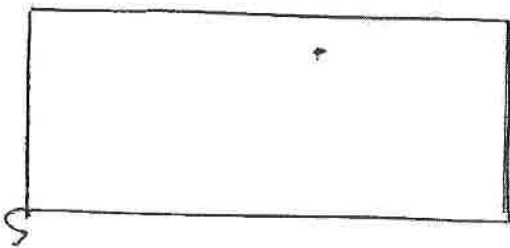


$P(A \cup B)$

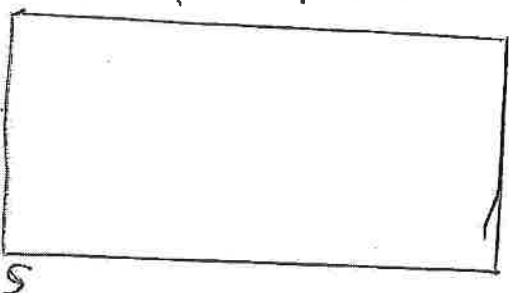
[A and B NOT MUTUALLY EXCLUSIVE]



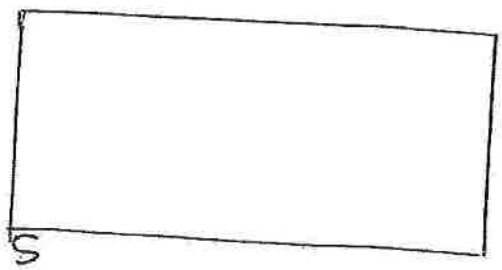
[A and B ARE MUTUALLY EXCLUSIVE]



$P(A' \cap B)$



$P(A' \cap B')$



$P(A \cup B')$

Probability

Exercise 1

- Write down the appropriate sample space for each of the following random experiments.
1. A hand of 4 cards is dealt from an ordinary pack of 52 playing cards and a count is made of the number of aces in the hand.
 2. A litter of six kittens is to be examined to determine how many are female.
 3. A box contains 3 white balls and 1 red ball. The balls are to be drawn from the box one after another until the red ball is drawn. A note is to be made of the number of balls drawn.
 4. Two balls are drawn one after the other from a bag which contains several red balls and several white balls. A record is to be made of the colours of the balls in the order in which they are drawn.
 5. A coin is tossed three times. (a) A note is to be made of the outcomes of the three tosses. (b) A note is to be made of the number of heads tossed.
 6. A coin is tossed until a head is obtained. A note is to be made of the number of times the coin has to be tossed

1.2 Events

Definition. An event is a property associated with the outcomes of a random experiment, or, in set theory terminology, a subset of the sample space.

For example, when a cubical die is thrown, events of possible interest may be (a) that the score obtained is an even number, (b) that the score obtained is a multiple of 3.

An event will be denoted by a capital letter from the early part of the alphabet (A, B, C, ...).

Example 1

A card is dealt from an ordinary pack of playing cards and its suit is to be noted. The sample space is $S = \{\text{Hearts, Clubs, Diamonds, Spades}\}$.

Examples of events in this case are

$A = \{\text{the dealt card will be red} = \{\text{Hearts, Diamonds}\},$

$B = \{\text{the dealt card will not be a Heart} = \{\text{Clubs, Diamonds, Spades}\}.$

Example 2

Two dice are thrown together and the sum of the scores is to be recorded. The sample space is

$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Examples of events associated with this experiment are

$A = \{\text{the sum of the scores will be a perfect square} = \{4, 9\},$

$B = \{\text{the sum of the scores will be an even number} = \{2, 4, 6, 8, 10, 12\}.$

Example 3

A coin is tossed three times and the outcomes are to be recorded in the order that they occur.

The sample space is

$S = \{(\text{HHH}), (\text{HHT}), (\text{HTH}), (\text{THH}), (\text{HTT}), (\text{THT}), (\text{TTH}), (\text{TTT})\},$

where, for example, the element (TTH) corresponds to the outcome that the first toss gives tail, the second gives head and the third gives head. Some examples of events in this case are:

$A = \{\text{exactly two heads will be tossed} = \{(\text{HHT}), (\text{HTH}), (\text{THH})\}$

$B = \{\text{the third toss will give a head} = \{(\text{HHH}), (\text{HTH}), (\text{TTH}), (\text{TTH})\}$

1.1 Random Experiments

Probability, as a subject, has been developed to deal with situations in which there is uncertainty. One class of such situations is that where the uncertainty is associated with a random experiment, defined as follows.

Definition. A random experiment is a course of action whose outcome is not predictable and is such that repeated trials (performances) of the action may give different outcomes.

Some examples of simple random experiments are as follows;

1. Tossing a coin.
2. Throwing an ordinary cubical die.
3. Dealing two cards from an ordinary pack of playing cards.

Definition. The set of all the possible outcomes of a random experiment is called the sample space and will be denoted by S.

Example 1

When tossing a coin and noting the face that is uppermost, the outcome may be a head (H) or a tail (T). In this case $S = \{\text{H}, \text{T}\}$, the curly brackets indicating that the order of the elements in the brackets is irrelevant, so that we could also write $S = \{\text{T}, \text{H}\}$.

Example 2

When throwing an ordinary cubical die and noting the score that is obtained, the outcome could be any one of 1, 2, 3, 4, 5, 6. Thus,

$S = \{1, 2, 3, 4, 5, 6\}$

Example 3

When dealing a card from an ordinary pack of 52 playing cards and noting its suit, the outcome could be any one of Hearts, Clubs, Diamonds or Spades. In this case

$S = \{\text{Hearts, Clubs, Diamonds, Spades}\}.$

Example 4

Consider the experiment when a cubical die is thrown twice. If a note is made of the individual scores obtained, then the sample space can be written as

$S = \{(x, y): x = 1, 2, 3, 4, 5, 6, y = 1, 2, 3, 4, 5, 6\},$

which consists of 36 elements, each being an ordered pair of numbers, the first (x) being the score on the first throw and the second (y) the score on the second throw. If, instead, a note is to be made of the sum of the two scores obtained, then the sample space will be

$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

Definition. The complement of an event A , denoted by A' , is the event that " A will not occur". The elements of A' will consist of all the elements of S that are not in A . For example, with A and B as defined in Example 2 above, we have
 $A' =$ the sum of the scores will not be a perfect square $= \{2, 3, 5, 6, 7, 8, 10, 11, 12\}$,
 $B' =$ the sum of the scores will not be an even number $= \{3, 5, 7, 9, 11\}$ (i.e. an odd number).

Combined events

Definitions. For any two events A and B

- the event that " A or B or both occur" is called the *union* of A and B and will be written as $A \cup B$;
- the event that " A and B will occur" is called the *intersection* of A and B and will be written as $A \cap B$.

With A and B as defined in Example 2 we have

$A \cup B =$ the sum of the scores will be a perfect square or an even number
 $= \{2, 4, 6, 8, 9, 10, 12\}$,

obtained by pooling the elements of A and the elements of B .

$A \cap B =$ the sum of the scores will be a perfect square and an even number $= \{4\}$,

obtained by listing those elements that are common to both A and B .

Example 4

For the events defined in Example 3 above, give a verbal description and list the elements of the events (a) $A \cup B$, (b) $A \cap C$, (c) $A \cap B'$.

From Example 3 we have

$S = \{(HHH), (HHT), (HTH), (THH), (HTT), (THT), (TTH), (TTT)\}$

$A =$ two heads will be tossed $= \{(HHT), (HTH), (THH)\}$

$B =$ the third toss will be a head $= \{(HHH), (HTH), (THH), (TTH)\}$

$C =$ the three tosses will have the same outcome $= \{(HHH), (TTT)\}$.

- $A \cup B =$ two heads will be tossed or the third toss will be a head
 $= \{(HHT), (HTH), (THH), (HHT), (TTH)\}$,

obtained by pooling the elements of A and B (i.e. by writing down the elements of A and then adding those elements of B that are not in A),

- $A \cap C =$ two heads will be tossed and the three tosses will have the same outcome.

We see that A and C have no element in common, so that $A \cap C$ cannot occur.

- $A \cap B' =$ two heads will be tossed and the third toss will not be a head.

$B' =$ the third toss will not be a head $= \{(HHT), (HTT), (THT), (TTT)\}$,

obtained by listing the elements of S that are not in B .

It follows that

$A \cap B' = \{(HHT)\}$, since (HHT) is the only common element in A and B' .

Definition. Two events which cannot occur at the same time are said to be *mutually exclusive*.

This means that there is no element in S which will result in both events occurring. A and C in Example 4 above are mutually exclusive events.

The union and intersection notation extends to more than two events in an obvious fashion.

Thus, for three events A , B and C ,

- $A \cup B \cup C =$ at least one of A , B , C will occur,

whose elements are obtained by pooling the elements of all three events,

- $A \cap B \cap C =$ all three events will occur,

whose elements will consist of all those elements which are common to A , B and C .

Example 5

With A , B and C as defined in Example 4 above, list the elements of $(A \cup B) \cap C$. In Example 4 we showed that

$A \cup B = \{(HHT), (HTH), (THH), (HHH), (TTH)\}$,

and $C = \{(HHH), (TTT)\}$.

Picking out the common elements we have $(A \cup B) \cap C = \{(HHH)\}$.

Exercise 1a2

- An ordinary cubical die is thrown. Let A denote the event that the score will be even, B the event that the score will be less than 3, and C the event that the score will be a multiple of 3.

- List the elements of S , A , B and C . State which two events are mutually exclusive.

- Give verbal descriptions and list the elements of the events

(i) A' , (ii) $A \cap B$, (iii) $A \cap C'$, (iv) $A \cup B \cup C$, (v) $(A \cup B) \cap C$.

- A cubical die is thrown twice. Write down the sample space for the outcome, expressing each element as an ordered pair. Determine the elements of each of the events

- $A =$ the sum of the scores will be divisible by 4,

- $B =$ the scores will be the same,

- $C =$ both scores will be even,

- $D =$ the scores will differ by at least 4,

- $A \cap C$, (f) $B \cup D$, (g) $B \cap A'$.

Which pairs (if any) of the events A , B , C and D are mutually exclusive?

- Two cards are dealt from a pack of 5 cards which are numbered from 1 to 5, and the sum of the two numbers dealt is recorded. Write down the 7 elements of the sample space and determine the elements of the events

- $A =$ the sum will be greater than 7,

- $B =$ the sum will be a prime number,

- $C =$ the sum will be an odd number,

- $A \cup C$, (e) $B \cap C$, (f) $A \cap B'$, (g) $A \cup B'$, (h) $A \cap B \cap C$.

Probability

The probability of an event occurring is the chance or likelihood of it occurring. The probability of an event A, written $P(A)$, can be between zero and one, with $P(A) = 1$ indicating that the event will certainly happen and with $P(A) = 0$ indicating that event A will certainly not happen.

$$\text{Probability} = \frac{\text{the number of successful outcomes of an experiment}}{\text{the number of possible outcomes}}$$

So, for example, if a coin were tossed, the probability of obtaining a head = $\frac{1}{2}$, since there are 2 possible outcomes (heads or tails) and 1 of these is the 'successful' outcome.

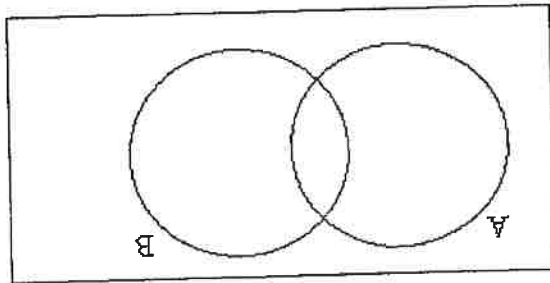
Using Set Notation

Probability can be studied in conjunction with set theory, with Venn Diagrams being particularly useful in analysis.
The probability of a certain event occurring, for example, can be represented by $P(A)$. The probability of a different event occurring can be written $P(B)$. Clearly, therefore, for two events A and B,

$$P(A) + P(B) - P(A \cap B) = P(A \cup B)$$

$P(A \cap B)$ represents the probability of A AND B occurring. $P(A \cup B)$ represents the probability of A OR B occurring.

This can be shown on a Venn diagram. The rectangle represents the sample space, which is all of the possible outcomes of the experiment. The circle labelled as A represents event A. In other words, all of the points within A represent possible ways of achieving the outcome of A. Similarly for B.



So in the diagram, $P(A) + P(B)$ is the whole of A (the whole circle) + the whole of B (so we have counted the middle bit twice).

A' is the complement of A and means everything not in A. So $P(A')$ is the probability that A does not occur. Note that the probability that A occurs + the probability that A does not occur = 1 (one or the other must happen). So $P(A) + P(A') = 1$. Thus:

$$P(A') = 1 - P(A)$$

Mutually Exclusive Events

Events A and B are **mutually exclusive** if they have no events in common. In other words, if A occurs B cannot occur and vice-versa. On a Venn diagram, this would mean that the circles representing events A and B would not overlap.

If, for example, we are asked to pick a card from a pack of 52, the probability that the card is red is $\frac{1}{2}$. The probability that the card is a club is $\frac{1}{4}$. However, if the card is red it can't be a club. These events are therefore mutually exclusive.

If two events are mutually exclusive, $P(A \cap B) = 0$, so

$$\bullet \quad P(A) + P(B) = P(A \cup B)$$

Independent Events

Two events are **independent** if the first one does not influence the second. For example, if a bag contains 2 blue balls and 2 red balls and two balls are selected randomly, the events are:

- a) independent if the first ball is replaced after being selected
- b) not independent if the first ball is removed without being replaced. In this instance, there are only three balls remaining in the bag so the probabilities of selecting the various colours have changed.

Two events are independent if (and only if):

$$\bullet \quad P(A \cap B) = P(A)P(B)$$

This is known as the multiplication law.

Conditional Probability

Conditional probability is the probability of an event occurring, given that another event has occurred. For example, the probability of John doing mathematics at A-level, given that he is doing physics may be quite high. $P(A|B)$ means the probability of A occurring, given that B has occurred. For two events A and B,

$$\bullet \quad P(A \cap B) = P(A|B)P(B)$$

and similarly $P(A \cap B) = P(B|A)P(A)$.

If two events are mutually exclusive, then $P(A|B) = 0$.

Independence

Using the above condition for independence, we deduce that if two events are independent, then:

$$P(A)P(B) = P(A|B)P(B) = P(B|A)P(A), \text{ or:}$$

$$P(A) = P(A|B) \text{ and } P(B) = P(B|A)$$

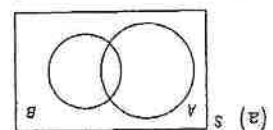
Example

A six-sided die is thrown. What is the probability that the number thrown is prime, given that it is odd.

The probability of obtaining an odd number is $3/6 = 1/2$. Of these odd numbers, 2 of them are prime (3 and 5).

$$P(\text{prime} | \text{odd}) = \frac{P(\text{prime and odd})}{P(\text{odd})} = \frac{2/6}{3/6} = 2/3$$

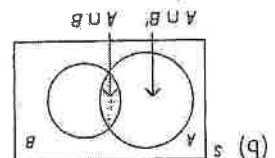
Other useful results relating two events A and B



$$P(A \cap B) = P(B \cap A)$$

$\downarrow$

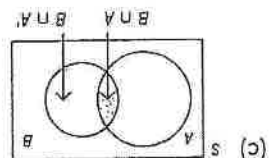
$$P(A \text{ and } B) \quad P(B \text{ and } A)$$



$$P(A) = P(A \cap B) + P(A \cap B')$$

$\downarrow$

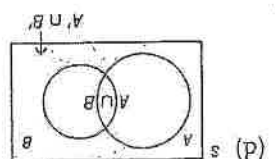
$$P(A \text{ and } B) \quad P(A \text{ but not } B)$$



$$P(B) = P(B \cap A) + P(B \cap A')$$

$\downarrow$

$$P(B \text{ and } A) \quad P(B \text{ but not } A)$$



$$P(\text{neither } A \text{ nor } B) = 1 - P(A \cup B)$$

$\downarrow$

$$\text{i.e.} \quad P(A' \cap B') = 1 - P(A \cup B)$$

Example

In a survey, 15% of the participants said that they had never bought lottery tickets or a premium bonds, 73% had bought lottery tickets and 49% had bought premium bonds.

Find the probability that a person chosen at random from those taking part in the survey

- (a) had bought lottery tickets or premium bonds,
- (b) had bought lottery tickets and premium bonds,
- (c) had bought lottery tickets only.

Solution

L: person has bought lottery tickets, $P(L) = 0.73$.

B: person has bought premium bonds, $P(B) = 0.49$.

$P(\text{neither } L \text{ nor } B) = 0.15$

(a) $P(L \text{ or } B) = 1 - P(\text{neither } L \text{ nor } B)$

$$= 1 - 0.15$$

$$= 0.85$$

(b) Use $P(L \text{ or } B) = P(L) + P(B) - P(L \text{ and } B)$

$$0.85 = 0.73 + 0.49 - P(L \text{ and } B)$$

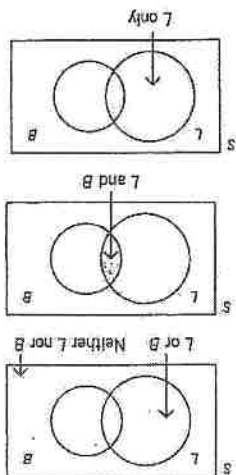
$$P(L \text{ and } B) = 0.73 + 0.49 - 0.85$$

$$= 0.37$$

(c) $P(L \text{ only}) = P(L) - P(L \text{ and } B)$

$$= 0.73 - 0.37$$

$$= 0.36$$



PROBABILITY

ADDITION RULE

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
or, using sets,
 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

If A and B are mutually exclusive events then $P(A \text{ and } B) = 0$,
i.e. $A \cap B$ is empty, then the addition rule becomes

$$P(A \text{ or } B) = P(A) + P(B)$$

MULTIPLICATION RULE

$$P(A \text{ and } B) = P(A) \cdot P(B/A)$$

$$P(A \cup B) = P(A) \cdot P(B/A)$$

or

If A and B are independent events then $P(B/A) = P(B)$ and the
multiplication rule becomes

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

or

$$P(A \cup B) = P(A) \cdot P(B)$$

EXAMPLE 1

A bag contains 3 yellow, 4 green and 5 red marbles. Three are drawn at random (i.e. without replacement).

Calculate the probability that

- (a) all are green
- (b) all are the same colour
- (c) two are red and one is green
- (d) they are all different colours
- (e) at least two are green.

SOLUTION

$$(a) \quad P(\text{g and g/g and g/g/g}) = 4/12 \times 3/11 \times 2/10$$

$$= 1/55.$$

$$(b) \quad P(\text{ggg or rrr or yyy}) = 1/55 + 5/12 \times 4/11 \times 3/10 + 3/12 \times 2/11 \times 1/10$$

$$= 3/44.$$

$$(c) \quad P(\text{RRG or RGR or GRR}) = 5/12 \times 4/11 \times 4/10 \times 3 \text{ (order)}$$

$$= 2/11.$$

$$(d) \quad P(\text{all different}) = P(\text{RGY}) \times 3! \text{ (order)}$$

$$= 5/12 \times 4/11 \times 3/10 \times 6$$

$$= 3/11.$$

$$(e) \quad P(\text{at least two green}) = P(2 \text{ green and 1 not or 3 green})$$

$$= 4/12 \times 3/11 \times 8/10 \times 3 + 1/55$$

$$= 13/55.$$

EXAMPLE 2

Two events A and B are such that $P(A) = 1/3$, $P(B) = 1/4$, $P(A \cap B) = 1/24$.

Evaluate

- $P(A/B)$
- $P(A \cup B)$
- $P(A' \cap B')$

Another event C is such that A and C are mutually exclusive and B and C are independent.

If $P(C) = 1/2$, calculate $P(A \cup B \cup C)$.

SOLUTION

(a) By the multiplication rule $P(A \cap B) = P(B) \cdot P(A/B)$ therefore

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{1/24}{1/4}$$

$$= 1/6$$

(b) We now have $P(B \cap A') = 5/24$

and therefore $P(A \cup B) = 1/3 + 5/24$

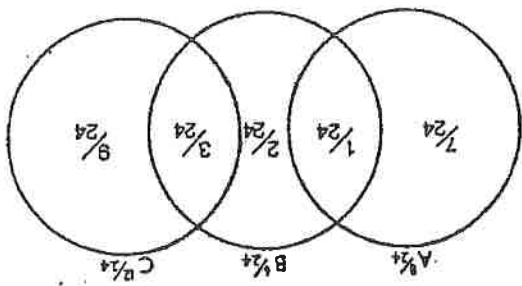
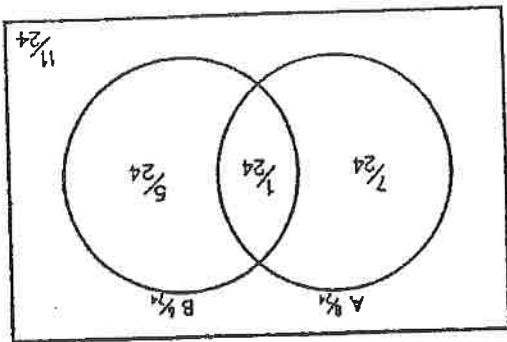
$$= 13/24$$

(c) $P(A' \cap B') = 1 - P(A \cup B)$

$$= 11/24$$

A and C being mutually exclusive means that $A \cap C = \emptyset$, see diagram.

Since B and C are independent $P(B \cap C) = P(B) \cdot P(C) = 1/8$.
 $P(A \cup B \cup C) = P(A) + P(B \cap A') + P(C \cap B') = 22/24$
 or $11/12$, see diagram.



EXAMPLE 3

The 28 children in a class are selected at random to play either soccer (10 pupils), badminton (4 pupils) or netball (14 pupils). Three of the pupils are called Anne, Bert and Charles.

Calculate the probabilities that

- (a) Anne plays netball,
- (b) Bert plays netball,
- (c) both Anne and Bert play netball,
- (d) all three play netball,
- (e) they all play the same sport,
- (f) they all play different sports,
- (g) two of them play the same sport and the third a different one.

SOLUTION

- (a) $P(\text{Anne plays netball}) = 14/28$
- $= 1/2$
- (b) $P(\text{Bert plays netball}) = 1/2$
- (c) $P(\text{A and B both play netball}) = 14/28 \times 13/27$
- $= 13/54$
- (d) $P(\text{all three play netball}) = 14/28 \times 13/27 \times 12/26$
- $= 1/9$
- (e) $P(\text{all the same}) = 1/9 + 4/28 \times 3/27 \times 2/26 + 10/28 \times 9/27 \times 8/26$
- $= 122/819$
- (f) $P(\text{all different}) = P(\text{A soccer, B badminton, C netball}) \cdot 3!$
- $= 10/28 \times 4/27 \times 14/26 \times 6$
- $= 20/117$
- (g) $P(2 \text{ play one sport, the third another}) = 1 - P(\text{all the same or all different})$
- $= 557/819$

EXAMPLE 4

Two players, A and B, take turns to throw a fair tetrahedral die numbered 1, 2, 3, 4. The player who first scores a 4 on the base wins. Calculate the probability that A wins and the probability that B wins.

SOLUTION

1st round	$1/4$	$3/4 \times 1/4$
2nd round	$(3/4)^2 \cdot 1/4$	$(3/4)^3 \cdot 1/4$
3rd round	$(3/4)^4 \cdot 1/4$	$(3/4)^5 \cdot 1/4$

A

B

...

$$P(A \text{ wins}) = 1/4 + (3/4)^2 \cdot 1/4 + (3/4)^4 \cdot 1/4 + \dots \quad (G.P. a = 1/4, r = (3/4)^2)$$

$$= \frac{1/4}{1 - (3/4)^2}$$

$$= \frac{4/7}{3/7} = \frac{4}{3}$$

$$P(B \text{ wins}) = \frac{3}{7}$$

EXERCISES

1. A box contains 9 discs of which 4 are red, 3 are white and 2 are blue. Three discs are to be drawn at random, without replacement.
Calculate the probabilities that: -
 (a) the discs, in order, will be coloured red, white and blue respectively
 (b) one disc of each colour will be drawn
 (c) the third disc will be red
 (d) no red disc will be drawn.
2. The probability that a hen lays an egg on a given day is 0.7. If it laid the previous day and 0.9 if it did not.
Calculate the probability that a hen lays on Tuesday given that it did lay on the previous Sunday.
3. In a class of 30 pupils, 5 have red hair and 12 study Russian. If a quarter of those who study Russian have red hair, find the probabilities that: -
 (a) a member of the class has red hair and studies Russian,
 (b) a child who studies Russian does not have red hair,
 (c) a member of the class neither studies Russian nor has red hair.
4. Broad bean seeds are planted in pairs because, on average only 70% of them germinate. A bed is planted with 100 pairs of seeds.
Calculate the expected number of pairs in which: -
 (a) both seeds germinate,
 (b) only one seed germinates,
 (c) no seeds germinate,
 (d) at least one seed germinates.
5. A driving instructor assesses that three of his pupils have probabilities of 5/6, 4/5 and 3/4 of passing their tests.
Calculate the probabilities that: -
 (a) exactly one pupil passes,
 (b) not more than one passes,
 (c) at least one passes.

6. A bag contains 6 black balls and 4 white balls. Balls are drawn at random without replacement. Calculate the probability that the balls alternate BWBW etc. until only black balls remain. Repeat to obtain a general formula if there are b black balls and w white ones initially.
7. In a class of 24 pupils, what is the probability of at least one repeated birthday. (Ignore leap years.)
8. A woman has three possible routes to drive to work, A, B and C, which she uses with respective probabilities $1/2$, $1/4$ and $1/4$. Using these routes gives her probabilities of being late of $1/12$, $1/4$ and $1/3$ respectively.

Calculate:-

- (a) the probability that she is late on any particular morning,
- (b) the probability that she travelled by route B on a morning she was late.

9. Two independent events A and B have associated probabilities of 3p and 4p of occurring.

If $P(A \cap B) = p$, calculate the value of p and find:-

- (a) $P(A/B)$,
- (b) $P(B/A)$,
- (c) $P(A \cap B)$.

A third event C is such that $P(A \cup C) = 5/8$ and $P(A \cap B \cap C) = 1/8$. If A and C are mutually exclusive, show that B and C are independent.

10. Three fair dice are thrown. Calculate the probabilities that:-

- (a) the total score is 17,
- (b) the total score is less than 17,
- (c) none of the individual scores is less than 2,
- (d) the lowest individual score is 2.

Exercise

1. One card is dealt from a shuffled pack of ordinary playing cards. It is replaced in the pack which is then shuffled and a card is dealt from it. Calculate the probabilities that (a) the first card is an ace, (b) at least one of the two cards dealt is an ace. [Compare your answers with those obtained in Example 3 above].
2. A fair cubical die has two of its faces numbered 1, two numbered 2 and two numbered 3. When the die is thrown the score obtained is the number on the uppermost face. Given that the die is thrown three times, calculate the probability that the three scores obtained (a) will be equal, (b) will sum to 6.
3. Stating any assumptions that you make, find an expression for the probability that in a group of n people at least two of them share the same birthmonth. Deduce the smallest n for which the group is more likely than not to include at least two people sharing the same birthmonth.
4. Three fair cubical dice are tossed together. Find the probability that the three scores will be such that (a) they are all different (b) they are three consecutive integers.
5. Two balls are drawn at random, one after another, from a bag which contains 10 red, 30 white and 20 blue balls. Calculate the probability that (a) both balls will be white, (b) the first ball will be white and the second blue, (c) at least one blue ball will be drawn.
6. Repeat Question 5, if the first ball drawn is replaced in the bag before the second ball is drawn.
7. Two cards are to be drawn at random from a pack of 15 cards which are numbered from 1 to 15, respectively. For each of the cases when the drawing is made (i) with replacement, and (ii) without replacement, calculate the probability that (a) the first number drawn will be less than 10 and the second drawn will be greater than 9, (b) the sum of the two numbers drawn will be 8.
8. A tetrahedral die has its four faces marked with the numbers 1, 2, 3 and 4, respectively. When the die is thrown onto a hard surface the score obtained is the number on the face in contact with the surface, which is equally likely to be any one of the four faces. The die is thrown twice. Calculate the probability that
 - (a) the sum of the two scores will be 5,
 - (b) the difference between the two scores will be 1,
 - (c) the product of the two scores will be 4.

Exercise

1. A and B are two events such that $P(A) = 0.7$, $P(B) = 0.2$, and $P(A \cap B) = 0.1$.
Evaluate (a) $P(A)$, (b) $P(A \cup B)$, (c) $P(A \cap B)$.
A and B are mutually exclusive events such that $P(A) = 0.5$ and $P(B) = 0.1$.
Evaluate (a) $P(A \cup B)$, (b) $P(A \cap B)$, (c) $P(A \cup B)$.
The events A and B are such that $P(A) = 0.4$, $P(B) = 0.3$, and $P(A \cup B) = 0.58$.
Evaluate (a) $P(A \cap B)$, (b) $P(A \cap B)$, (c) $P(A \cup B)$.
A and B are two events such that $P(A) = 0.25$, $P(B) = 0.40$ and $P(A \cap B) = 0.15$.
Find (a) $P(A)$, (b) $P(A \cup B)$, (c) $P(A \cap B)$, (d) $P(A \cap B)$, (e) $P(A \cap B)$.
A and B are two events such that $P(A) = 0.36$, $P(B) = 0.25$, $P(A \cap B) = 0.24$.
Evaluate (a) $P(A)$, (b) $P(A \cap B)$, (c) $P(A \cap B)$, (d) the probability that only one of A and B will occur.
2. A, B and C are three events with A and B being mutually exclusive and such that $P(A) = 0.4$, $P(B) = 0.3$, $P(C) = 0.2$, $P(B \cap C) = 0.1$, and $P(A \cap C) = 0.08$. By means of a Venn diagram, or otherwise, evaluate (a) $P(A \cap C)$, (b) $P(B \cap C)$, (c) $P(A \cup B \cup C)$, (d) the probability that none of the events A, B and C will occur.
3. The events A, B and C are such that $P(A) = \frac{1}{2}$, $P(B) = \frac{3}{4}$, $P(C) = \frac{1}{4}$ and $P(A \cap B) = \frac{1}{6}$.
Calculate (a) $P(A \cup B)$, (b) $P(A \cap B)$, (c) $P(A \cap B)$.
(d) Given that A and C are mutually exclusive and that B and C are mutually exclusive, calculate $P(A \cup B \cup C)$.

ie 8 Events

INDEPENDENT

CONDITIONAL PROBABILITY

Given that $P(A) = 0.4$ and $P(A \cup B) = 0.7$, determine the value of $P(B)$ in each of the cases when

- (i) A and B are mutually exclusive.
- (ii) A and B are independent.

[3]

Two events A and B are such that

$$P(A) = 0.5, P(B) = 0.3, \text{ and } P(A \cap B) = 0.1.$$

Calculate (i) $P(A \cup B)$, (ii) $P(A \cap B)$, (iii) the probability that exactly one of the two events occurs.

(2)

The two events A and B are such that

$$P(A) = 0.2, P(B) = 0.4, \text{ and } P(A|B) = 0.3.$$

(i) State, with your reason, whether or not A and B are independent.

(ii) Find the values of $P(A \cup B)$ and $P(B|A)$.

[6]

Two events A and B are such that $P(A) = 0.6$, $P(A \cap B) = 0.2$ and $P(A \cup B) = 0.7$. Calculate

- (i) $P(B)$,
- (ii) $P(B|A)$.

(4)

The three events A , B and C , are such that

$$P(A) = \frac{1}{5}, P(B) = \frac{1}{10}, P(A \cup B) = \frac{15}{7}, P(B \cup C) = \frac{23}{60}$$

and the events A and C are independent.

- (i) Show that $P(C) = \frac{3}{5}$.
- (ii) Determine whether or not B and C are independent.
- (iii) Given further that A and B are mutually exclusive, evaluate $P(A \cup B|C)$.

[8]

The two events A and B are independent and their probabilities of occurring are $P(A) = \frac{2}{3}$ and $P(B) = \frac{5}{7}$.

- (i) Show that $P(A \cup B) = \frac{4}{3}$.
- (ii) Calculate the conditional probability that only one of the two events will occur.
- (iii) Calculate the conditional probability that A occurred given that only one of the two events occurred.

Another event C is such that A and C are mutually exclusive, $P(C) = \frac{1}{5}$, and $P(B' \cap C) = \frac{15}{7}$.

- (iv) $P(B|A \cup C)$,
- (v) $P(B|A \cup C)$.

Calculate

(6)

DISCRETE PROBABILITY DISTRIBUTIONS

DEFINITION

X is a discrete random variable if it can take discrete values $x_1, x_2, x_3, \dots, x_n$ with associated probabilities $p_1, p_2, p_3, \dots, p_n$

such that

$$\sum_{i=1}^n p_i = 1$$

MEAN

The theoretical mean value of X, or the EXPECTATION of X, is given by

$$E(X) = \sum x \cdot P(X=x)$$

VARIANCE

$$\begin{aligned} \text{Var}(X) &= \sum x^2 \cdot P(X=x) - \text{mean}^2 \\ &= \sum x^2 P(X=x) - \left[\sum x \cdot P(X=x) \right]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

EXPECTATION AND VARIANCE OF OTHER FUNCTIONS

$$1 \quad E\{f(X)\} = \sum f(x) \cdot P(X=x) \text{ where } f(X) \text{ is any function of } X.$$

$$2 \quad E(X+c) = E(X) + c \quad \text{and} \quad \text{Var}(X+c) = \text{Var}(X) \text{ where } c \text{ is a constant.}$$

$$3 \quad E(cX) = c \cdot E(X) \quad \text{and} \quad \text{Var}(cX) = c^2 \cdot \text{Var}(X) \text{ where } c \text{ is a constant.}$$

$$4 \quad E(X+Y) = E(X) + E(Y) \quad \text{and} \quad \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) \text{ if } X, Y \text{ independent.}$$

$$5 \quad E(X-Y) = E(X) - E(Y) \quad \text{BUT} \quad \text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) \text{ if } X, Y \text{ independent.}$$

$$6. \quad E(ax+b) = aE(X) + b$$

$$7. \quad \text{Var}(ax+b) = a^2 \text{Var}(X)$$

11/4/

EXAMPLE 1

For the following probability distribution calculate

(a) $E(X)$, (b) $E(X^2)$, (c) $\text{Var}(X)$.

X	4	6	8	10	12
$P(X)$	$1/8$	$1/6$	$3/8$	$1/4$	$1/12$

SOLUTION

$$\begin{aligned} \text{(a) } E(X) &= 4 \times 1/8 + 6 \times 1/6 + 8 \times 3/8 + 10 \times 1/4 + 12 \times 1/12 = \bar{8} \\ \text{(b) } E(X^2) &= 4^2 \times 1/8 + 6^2 \times 1/6 + 8^2 \times 3/8 + 10^2 \times 1/4 + 12^2 \times 1/12 \\ &= \bar{69} \end{aligned}$$

$$\text{(c) } \text{Var}(X) = E(X^2) - E^2(X)$$

$$= 69 - 8^2$$

$$= \bar{5}$$

EXAMPLE 2

The number of eggs X laid by a lesser spotted willow warbler can take the values 0, 1, 2, 3, 4 with probabilities proportional to X . Find these probabilities. Calculate the mean and variance for this probability distribution. Eggs are collected from three nests. Calculate the probability that only four eggs are collected altogether.

SOLUTION

Since the probabilities are proportional to X and their sum is unity, the probability distribution is as follows: -

X	0	1	2	3	4
$P(X)$	0	$1/10$	$2/10$	$3/10$	$4/10$

$$E(X) = 1 \times 1/10 + 2 \times 2/10 + 3 \times 3/10 + 4 \times 4/10 = 3$$

$$\text{Var}(X) = 1^2 \times 1/10 + 2^2 \times 2/10 + 3^2 \times 3/10 + 4^2 \times 4/10 - 3^2 = 1$$

Since the probability that a nest contains 0 eggs is 0, the only possibilities giving a total of 4 eggs are 1, 1, 2; 1, 2, 1; 2, 1, 1. Therefore $P(4 \text{ eggs}) = 1/10 \times 1/10 \times 2/10 \times 3 = 3/500$

EXAMPLE 3

p is the constant probability that an event will occur in a single trial of an experiment and $q (= 1 - p)$ is the constant probability of failure.

If the experiment is repeated, successive trials being independent, until the first success occurs, derive an expression for the probability function $P(X)$, the probability that the first success occurs on the X^{th} trial.

Show that $EP(X) = 1$ and show that the mean value of X is $1/p$.

A single card is drawn from a well shuffled pack of 52. What is the probability that it is a spade?

This is repeated, with replacement, until a spade is drawn. How many trials, on average, will be needed until a spade is drawn?

What is the probability that this number of trials is necessary?

SOLUTION

$P(\text{1st success on } x^{\text{th}} \text{ trial}) = q \times q \times q \times \dots \times q \times p = q^{x-1} \cdot p$

$$\begin{aligned} EP(X) &= p + q \cdot p + q^2 \cdot p + q^3 \cdot p + \dots \\ &= p(1 + q + q^2 + q^3 + \dots) \\ &= p(1 - q)^{-1} \\ &= p/p \\ &= 1 \end{aligned}$$

$$\begin{aligned} E(X) &= 1 \times p + 2 \times q \cdot p + 3 \times q^2 \cdot p + 4 \times q^3 \cdot p + \dots \\ &= p(1 + 2q + 3q^2 + 4q^3 + \dots) \\ &= p(1 - q)^{-2} \\ &= p \times 1/p^2 \\ &= 1/p \\ P(\text{Spade}) &= 1/4 \end{aligned}$$

So we have the above distribution with $p = 1/4$ and the mean number of trials is 4.

And the probability of 4 trials being needed is $(3/4)^3 \times 1/4$ i.e. 27/256

EXAMPLE 4

Two tetrahedral dice are thrown together and the total score, X is recorded. Draw up a table for the probability distribution for X .

Calculate the mean and variance of the following functions: -

- (a) X , (b) $2X$, (c) $2X + 4$.

SOLUTION

X	2	3	4	5	6	7	8
$P(X)$	1/16	1/8	3/16	1/4	3/16	1/8	1/16

$$E(X) = 5 \text{ by symmetry.}$$

$$Var(X) = 2.5 \text{ see previous examples.}$$

$$E(2X) = 2 \times E(X) = 10$$

$$Var(2X) = 2^2 \times Var(X) = 10$$

$$E(2X + 4) = E(2X) + 4 = 14$$

$$Var(2X + 4) = Var(2X) = 10$$

EXERCISES

1. The number of eggs laid by giant tortoises have the following distribution: -

no. of eggs X	1	2	3	4	5
$P(X)$	1/10	2/10	3/10	3/10	1/10

Calculate (a) $E(X)$, (b) $E(X^2)$ (c) $\text{Var}(X)$.

2. The random variable X has a Binomial distribution with $n = 16$ and $p = 1/4$ and the random variable Y has a Poisson distribution with mean 1.

Find the mean and standard deviation of

- (a) $X - Y$ (b) $X + 3$ (c) $3X + 4Y$.

(Assume X and Y are independent. See later for Binomial and Poisson distributions.)

3. The random variable X can take values 1, 2, 3, 4, 5 with equal probabilities (rectangular distribution). Find $E(X)$ and $\text{Var}(X)$.
Repeat to obtain formulae for the mean and variance of the rectangular distribution when X can take values 1, 2, 3, ..., n .

4. A fair six sided die has its faces numbered as follows; one 1, two 2's, three 3's. If it is thrown once, calculate the mean and variance of the score. If X is the total score when the die is thrown twice, draw up a table to show the probability distribution for X and calculate the mean and variance of X .

5. A tetrahedral die is such that the face on the base, X , can take values 1, 2, 3, 4. If $P(X = 4)$ is twice as likely as each of the other scores, find $E(1/X)$.

6. X is the number of heads when three fair coins are thrown together and Y is the number on the base of a fair tetrahedral die with faces 1, 2, 3, 4. Draw up tables to show the probability distributions of X and Y. Calculate

(a) $P(X > Y)$, (b) $P(Y > X)$.

7. A fair die is thrown repeatedly until a six appears. If X denotes the number of throws before the first six appears, obtain the probability distribution for X and find $E(X)$ and $Var(X)$.

8. Rebecca has a bunch of 3 keys, one of which fits a lock. If she tries the keys at random (i.e. forgetting which she has tried previously) obtain the probability distribution for X, the number of attempts made until the lock opens. Julian has more sense and tries the keys in sequence. Obtain the distribution for Y, the number of attempts Julian makes. Find

(a) $P(X < Y)$, (b) $P(Y < X)$.

9. Adam and Betty each have a fair six sided die numbered 1, 2, 3, 4, 5, 6. Adam throws his die once and records X, his score multiplied by 3. Betty throws her die 3 times and records Y, the sum of the 3 scores. Calculate the mean and variance for a single throw of a die and hence deduce the mean and variance for both X and Y.

10. At a roll-a-ball stall at the funfair the stake is 10p for 4 balls. Each roll of a ball is independent of any others and the balls can roll into slots with scores as follows: -

Score X	P(X)
1	1/8
2	2/8
3	2/8
4	2/8
5	1/8

Prizes are awarded as follows: -

Total score	Prize
18 or more	£1
8	20p
less than 8	50p

Calculate the expected prize money from a roll of 4 balls.

120/

Discrete Random Variables

A probability distribution is a table of values showing the probabilities of various outcomes of an experiment.

For example, if a coin is tossed three times, the number of heads obtained can be 0, 1, 2 or 3. The probabilities of each of these possibilities can be tabulated as shown:

Number of Heads	0	1	2	3
Probability	1/8	3/8	3/8	1/8

A discrete variable is a variable which can only take a countable number of values. In this example, the number of heads can only take 4 values (0, 1, 2, 3) and so the variable is discrete. The variable is said to be random if the sum of the probabilities is one.

Probability Density Function

This term usually applies to continuous random variables, but is increasingly being applied to discrete random variables. Ideally for this case it should be called the probability function.

Properties of the probability (density) function.

$$P(X=x) \geq 0 \text{ and } \sum P(X=x) = 1$$

(p.d.f.) of X (or probability mass function) is a function which allocates probabilities. Put simply, it is a function which tells you the probability of certain events occurring. The usual notation that is used is

$$P(X=x) = \text{something.}$$

The random variable (r.v.) X is the event that we are considering. So in the above example, X represents the number of heads that we throw. So $P(X=0)$ means the probability that no heads are thrown. Here, $P(X=0) = 1/8$ (the probability that we throw no heads is 1/8).

In the above example, we could therefore have written:

x	0	1	2	3
$P(X=x)$	1/8	3/8	3/8	1/8

Quite often, the probability density function will be given to you in terms of x. In the above example, $P(X=x) = {}^3C_x (1/2)^3$ (see permutations and combinations for the meaning of ${}^3C_x$).

Example

A die is thrown repeatedly until a 6 is obtained. Find the probability density function for the number times we throw the die.

Let X be the random variable representing the number of times we throw the die.

$P(X = 1) = 1/6$ (if we only throw the die once, we get a 6 on our first throw. The probability of this is $1/6$).

$P(X = 2) = (5/6) \times (1/6)$ (if we throw the die twice before getting a 6, we must throw something that isn't a 6 with our first throw, the probability of which is $5/6$ and we must throw a 6 on our second throw, the probability of which is $1/6$)

etc

In general, $P(X = x) = (5/6)^{(x-1)} \times (1/6)$

Cumulative Distribution Function

The cumulative distribution function (c.d.f.) of a discrete random variable X is the function $F(t)$ which tells you the probability that X is less than or equal to t . So if X has p.d.f. $P(X = x)$, we have:

$$F(t) = P(X \leq x) = \sum P(X = x).$$

In other words, for each value that X can be which is less than or equal to t , work out the probability that X is that value and add up all such results.

Example

In the above example where the die is thrown repeatedly, let's work out $P(X \leq x)$ for some values of x .

$P(X \leq 1)$ is the probability that the number of throws until we get a 6 is less than or equal to 1. So it is either 0 or 1.

$$P(X = 0) = 0 \text{ and } P(X = 1) = 1/6. \text{ Hence } P(X \leq 1) = 1/6$$

$$\text{Similarly, } P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0 + 1/6 + 5/36 = 11/36$$

Expectation

The expected value (or mean) of X , where X is a discrete random variable, is a weighted average of the possible values that X can take, each value being weighted according to the probability of that event occurring. The expected value of X is usually written as $E(X)$ or m .

$$E(X) = \sum_{all\ x} x \cdot P(X = x)$$

So the expected value is the sum of: [(each of the possible outcomes) $\times$ (the probability of the outcome occurring)].

In more concrete terms, the expectation is what you would expect the outcome of an experiment to be on average.

Example

What is the expected value when we roll a fair die?

There are six possible outcomes: 1, 2, 3, 4, 5, 6. Each of these has a probability of $1/6$ of occurring. Let X represent the outcome of the experiment.

Therefore $P(X = 1) = 1/6$ (this means that the probability that the outcome of the experiment is 1 is $1/6$)

$P(X = 2) = 1/6$ (the probability that you throw a 2 is $1/6$)

$P(X = 3) = 1/6$ (the probability that you throw a 3 is $1/6$)

$P(X = 4) = 1/6$ (the probability that you throw a 4 is $1/6$)

$P(X = 5) = 1/6$ (the probability that you throw a 5 is $1/6$)

$P(X = 6) = 1/6$ (the probability that you throw a 6 is $1/6$)

$$E(X) = 1 \times P(X = 1) + 2 \times P(X = 2) + 3 \times P(X = 3) + 4 \times P(X = 4) + 5 \times P(X = 5) + 6 \times P(X = 6)$$

$$\text{Therefore } E(X) = 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = 7/2$$

So the expectation is 3.5. If you think about it, 3.5 is halfway between the possible values the die can take and so this is what you should have expected.

Expected Value of a Function of X

To find $E[f(X)]$, where $f(X)$ is a function of X , use the following formula:

$$\bullet E[f(X)] = \sum_{\text{all } x} f(x)P(X=x)$$

Example

For the above experiment (with the die), calculate $E(X^2)$

Using our notation above, $f(x) = x^2$

$$f(1) = 1, f(2) = 4, f(3) = 9, f(4) = 16, f(5) = 25, f(6) = 36$$
$$P(X=1) = 1/6, P(X=2) = 1/6, \text{ etc}$$

$$\text{So } E(X^2) = 1/6 + 4/6 + 9/6 + 16/6 + 25/6 + 36/6 = 91/6 = 15.167$$

The expected value of a constant is just the constant, so for example $E(1) = 1$. Multiplying a random variable by a constant multiplies the expected value by that constant, so $E[2X] = 2E[X]$.

A useful formula, where a and b are constants, is:

$$\bullet E[aX + b] = aE[X] + b$$

[This says that expectation is a linear operator].

Variance

The variance of a random variable tells us something about the spread of the possible values of the variable. For a discrete random variable X , the variance of X is written as $\text{Var}(X)$.

$$\bullet \text{Var}(X) = E[(X - m)^2]$$

where m is the expected value $E(X)$

This can also be written as:

$$\bullet \text{Var}(X) = E(X^2) - m^2$$

The *standard deviation* of X is the square root of $\text{Var}(X)$.

Note that the variance does not behave in the same way as expectation when we multiply and add constants to random variables. In fact:

- $\text{Var}[aX + b] = a^2 \text{Var}(X)$

You is because: $\text{Var}[aX + b] = E[(aX + b)^2] - (E[aX + b])^2$.

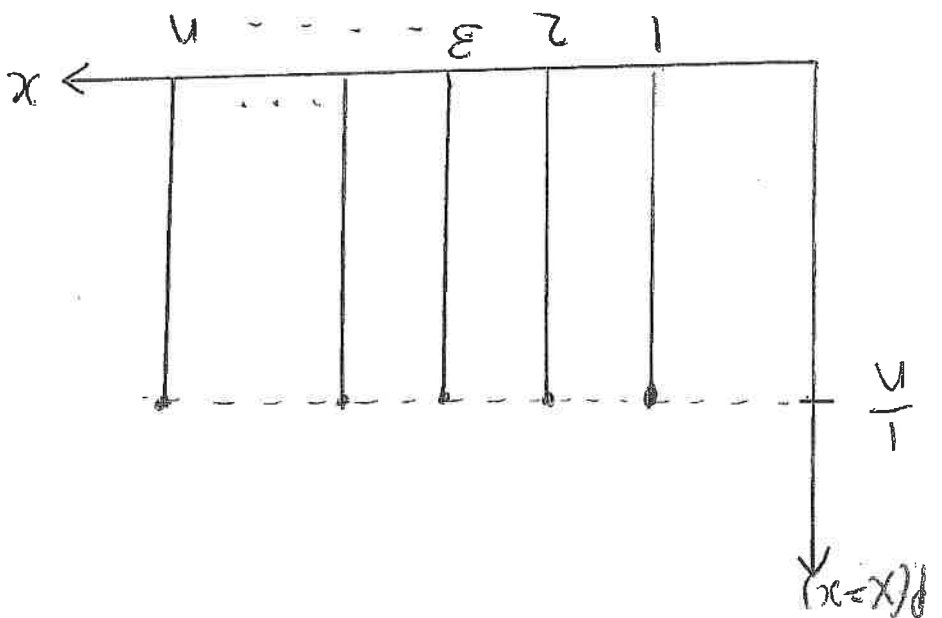
$$= E[a^2 X^2 + 2abX + b^2] - (aE(X) + b)^2$$

$$= a^2 E(X^2) + 2abE(X) + b^2 - a^2 E^2(X) - 2abE(X) - b^2$$

$$= a^2 E(X^2) - a^2 E^2(X) = a^2 \text{Var}(X)$$

Discrete Uniform Distribution:

A discrete uniform distribution is a symmetric probability distribution whereby a finite number of values are equally likely to be observed. That is, every one of 'n' possible values has an equal probability of $\frac{1}{n}$.



all outcomes equally likely to occur.

$$f(X=x) = \frac{1}{n} \text{ for } x=1, 2, 3, \dots, n$$

$$= 0 \text{ otherwise.}$$

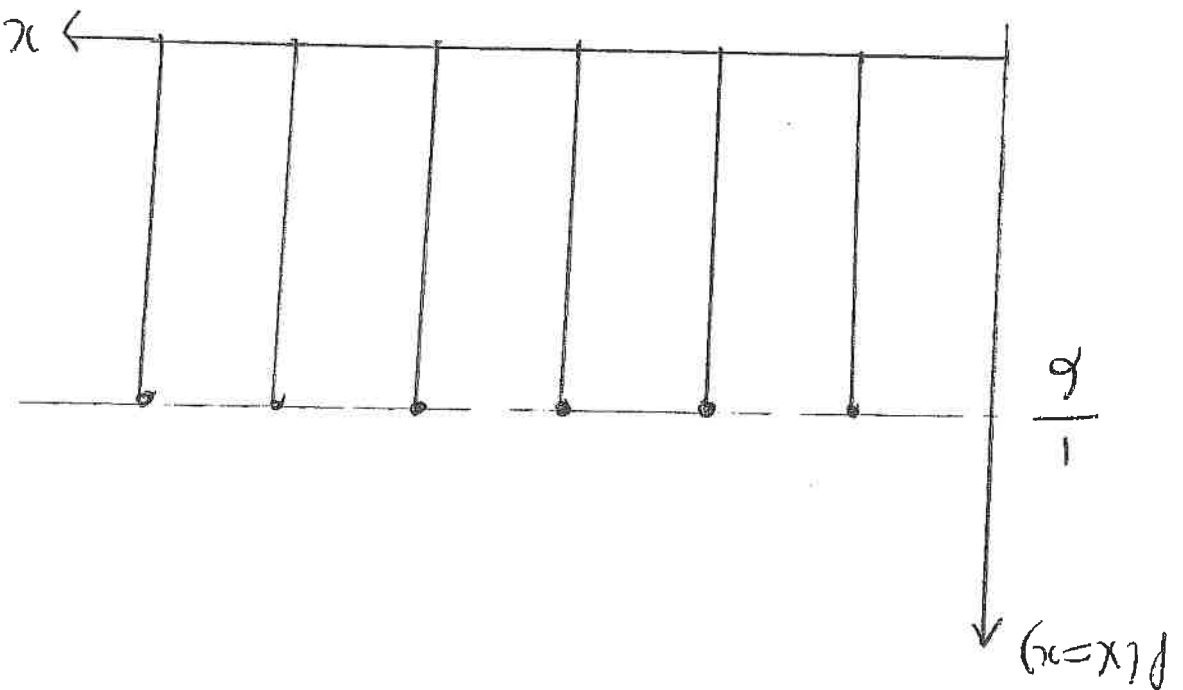
$$E(X) = \frac{n+1}{2}$$

$$\text{Var}(X) = \frac{n^2-1}{12}$$

Ex. Consider rolling a fair cubical die

$$S = \{1, 2, 3, 4, 5, 6\}$$

x	$P(X=x)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$



So $P(X=x) = \frac{1}{6}$ for $x = 1, 2, 3, 4, 5, 6$

otherwise $= 0$

12/1

Permutations and Combinations

Arranging Objects

- The number of ways of arranging n unlike objects in a line is $n!$ (pronounced 'n factorial'). $n! = n \times (n-1) \times (n-2) \times \dots \times 3 \times 2 \times 1$

Example

How many different ways can the letters P, Q, R, S be arranged?

The answer is $4! = 24$.

This is because there are four spaces to be filled: _ , _ , _ , _

The first space can be filled by any one of the four letters. The second space can be filled by any of the remaining 3 letters. The third space can be filled by any of the 2 remaining letters and the final space must be filled by the one remaining letter. The total number of possible arrangements is therefore $4 \times 3 \times 2 \times 1 = 4!$

- The number of ways of arranging n objects, of which p of one type are alike, q of a second type are alike, r of a third type are alike, etc is:

$$\frac{n!}{p! q! r! \dots}$$

Example

In how many ways can the letters in the word: STATISTICS be arranged?

There are 3 S's, 2 T's and 3 T's in this word, therefore, the number of ways of arranging the letters are:

$$10! = 50400$$

$$3! 2! 3!$$

Rings and Roundabouts

- The number of ways of arranging n unlike objects in a ring when clockwise and anticlockwise arrangements are different is $(n-1)!$

When clockwise and anti-clockwise arrangements are the same, the number of ways is $\frac{1}{2}(n-1)!$

Example

Ten people go to a party. How many different ways can they be seated?

Anti-clockwise and clockwise arrangements are the same. Therefore, the total number of ways is $\frac{1}{2}(10-1)! = 181\,440$

Combinations

The number of ways of selecting r objects from n unlike objects is:

$$\binom{n}{r} = {}^nC_r = \frac{n!}{r!(n-r)!}$$

Example

There are 10 balls in a bag numbered from 1 to 10. Three balls are selected at random. How many different ways are there of selecting the three balls?

$${}^{10}C_3 = \frac{10!}{3!7!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120$$

Permutations

A permutation is an ordered arrangement.

- The number of ordered arrangements of r objects taken from n unlike objects is:

$${}_nP_r = \frac{n!}{(n-r)!}$$

Example

In the Match of the Day's goal of the month competition, you had to pick the top 3 goals out of 10. Since the order is important, it is the permutation formula which we use.

$${}_{10}P_3 = \frac{10!}{7!}$$

$$= \frac{10!}{7!}$$

There are therefore 720 different ways of picking the top three goals.

Probability

The above facts can be used to help solve problems in

probability.

Example

In the National Lottery, 6 numbers are chosen from 49. You win if the 6 balls you pick match the six balls selected by the machine. What is the probability of winning the National Lottery?

The number of ways of choosing 6 numbers from 49 is ${}_{49}C_6 =$

$$13\,983\,816.$$

Therefore the probability of winning the lottery is $1/13\,983\,816 = 0.000\,000\,071\,5$ (3sf), which is about a 1 in 14 million chance.

The Binomial Distribution

If a discrete random variable X has the following probability density function (p.d.f.), it is said to have a binomial distribution:

$$\bullet \quad P(X = x) = {}^nC_x q^{(n-x)} p^x, \text{ where } q = 1 - p$$

p can be considered as the probability of a success, and q the probability of a failure.

Note: ${}^nC_r$ ("n choose r") is more commonly written $\binom{n}{r}$, but I shall use the former because it is easier to write on a computer. It means the number of ways of choosing r objects from a collection of n objects (see permutations and combinations).

If a random variable X has a binomial distribution, we write $X \sim B(n, p)$ ($\sim$ means 'has distribution...').

n and p are known as the parameters of the distribution (n can be any integer greater than 0 and p can be any number between 0 and 1). All random variables with a binomial distribution have the above p.d.f., but may have different parameters (different values for n and p).

Example

A coin is thrown 10 times. Find the probability density function for X , where X is the random variable representing the number of heads obtained.

The probability of throwing a head is $\frac{1}{2}$ and the probability of throwing a tail is $\frac{1}{2}$. Therefore, the probability of throwing 8 tails is $(\frac{1}{2})^8$

If we throw 2 heads and 8 tails, we could have thrown them HTTTTHTT, or TTHTTTT, or in a number of other ways. In fact, the total number of ways of throwing 2 heads and 8 tails is ${}^{10}C_2$ (see the permutations and combinations section).

Hence the probability of throwing 2 heads and 8 tails is ${}^{10}C_2 \times (\frac{1}{2})^2 \times (\frac{1}{2})^8$. As you can see this has a Binomial distribution, where $n = 10$, $p = \frac{1}{2}$.

You can see, therefore, that the p.d.f. is going to be: $P(X = x) = {}^{10}C_x (\frac{1}{2})^x (1 - \frac{1}{2})^{10-x}$.

From this, we can work out the probability of throwing, for example, 3 heads (put $x = 3$).

Expectation and Variance

If $X \sim B(n, p)$, then the expectation and variance is given by:

- $E(X) = np$
- $\text{Var}(X) = npq$

Example

In the above example, what is the expected number of heads thrown?

$$E(X) = np$$

Now in the above example, $p = \text{probability of throwing a head} = \frac{1}{2}$. $n = \text{number of throws} = 10$

Hence expected number of heads = 5.

This is what you would expect: if you throw a coin 10 times you would expect 5 heads and 5 tails on average.

THE BINOMIAL DISTRIBUTION

DEFINITION

If p is the constant probability that an event occurs in a single trial and $q (= 1 - p)$ is the probability that it does not occur then the probability of x successes in n independent trials is given by: -

$$P(X = x) = {}^nC_x \cdot p^x \cdot q^{n-x}$$

where x covers the sample space $0, 1, 2, 3, \dots, n$

MEAN AND VARIANCE

If $x \sim B(n, p)$ then $E(X) = np$ and $\text{Var}(X) = npq$.

EXAMPLE 1

If 20% of garments produced in a clothing factory are defective, find the probability that if 5 garments are chosen at random: -

(a) 1 is defective,

(b) 3 are defective,

(c) at most 3 are defective.

If 100 such samples of 5 were drawn, how many of them would be expected to have no defective items?

SOLUTION

Let X be the number of defectives in a sample of 5.

Then $X \sim B(5, 1/5)$.

(a) $P(X = 1)$

$$= 5 \times (1/5) \times (4/5)^4$$

$$= 0.410 \quad (\text{or using cumulative binomial tables, } 0.737 - 0.328)$$

$$= 51/3121 \times (1/5)^3 \times (4/5)^2$$

(b) $P(X = 3)$

$$= 0.051 \quad (0.993 - 0.942)$$

(c) $P(X = 0, 1, 2, \text{ or } 3)$

$$= 1 - P(X = 4 \text{ or } 5)$$

$$= 1 - 5 \times (1/5)^4 \times (4/5) - (1/5)^5$$

$$= 0.993 \quad (\text{or directly from tables})$$

$$= (4/5)^5$$

$$= 0.328 \quad (\text{or directly from tables})$$

$$= 33 \text{ (say)}$$

$$\text{So } 100 \times 0.328$$

samples would be expected to have no defectives.

EXAMPLE 2

Derive the mean and variance of the Binomial distribution.

If X has a Binomial distribution with mean 9 and variance 4.95 find the probability that X is greater than 10.

SOLUTION

$$E(X) = \sum x \cdot P(X = x)$$

$$= 0 \cdot q^n + 1 \cdot nq^{n-1}p + 2 \cdot \frac{n(n-1)}{2!} q^{n-2}p^2 + \dots + n \cdot p^n$$

$$+ 3 \cdot \frac{n(n-1)(n-2)}{3!} q^{n-3}p^3 + \dots + n \cdot p^n$$

$$= np(q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{2!} q^{n-3}p^2 + \dots + p^{n-1})$$

$$= np(q + p)^{n-1}$$

$$= np \text{ since } p + q = 1.$$

$$E(X^2) = \sum x^2 \cdot P(X = x)$$

$$= 0^2 q^n + 1^2 nq^{n-1}p + 2^2 \frac{n(n-1)}{2!} q^{n-2}p^2 + 3^2 \frac{n(n-1)(n-2)}{3!} q^{n-3}p^3 + \dots + n^2 p^n$$

$$= np\{q^{n-1} + 2(n-1)q^{n-2}p + 3 \frac{(n-1)(n-2)}{2!} q^{n-3}p^2 + \dots + np^{n-1}\}$$

$$= np\{[q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{2!} q^{n-3}p^2 + \dots + p^{n-1}]\}$$

$$+ [(n-1)q^{n-2}p + 2 \frac{(n-1)(n-2)}{2!} q^{n-3}p^2 + \dots + (n-1)p^{n-1}]\}$$

$$= np\{[q + p]^{n-1} + (n-1)p[q + p]^{n-2} + \dots + p^{n-2}\}$$

$$= np\{1 + (n-1)p[q + p]^{n-2}\}$$

$$= np\{1 + (n-1)p\}$$

$$= np(1 - p) + n^2 p^2$$

$$\text{Now Var}(X) = E(X^2) - E^2(X)$$

$$= np(1 - p) + n^2 p^2 - (np)^2$$

$$= npq$$

$$np = 9$$

$$npq = 4.95$$

$$\text{thus } q = 0.55$$

$$\text{and } p = 0.45$$

$$\text{and } n = 20$$

$$\text{i.e. } X \sim B(20, 0.45)$$

$$P(X > 10) = 1 - P(X \leq 10)$$

$$= 1 - 0.751$$

(tables)

$$= 0.249$$

EXAMPLE 3

Show that if $X \sim B(n, p)$ then $\frac{P(X = x)}{P(X = x - 1)} = \frac{(n - x + 1)p}{xq}$

(a) If $X \sim B(8, p)$ and $P(X = 5) = 6P(X = 6)$ find the value of p .

(b) Coffee tables have tiled tops consisting of 12 tiles.

The probability that a single tile is damaged is p .

If no tiles on a table top are damaged it is sold as perfect for £50.

If only one tile is damaged it is sold as a "second" for £30.

If more than one tile is damaged it is scrapped.

It is found that, of all tables sold, $1/3$ of them are seconds.

Calculate the value of p .

Calculate the proportion of tables which are scrapped.

The tables cost £20 each to make, calculate the expected profit from 100 tables.

SOLUTION

$$\frac{P(X = x)}{P(X = x - 1)} = \frac{n C_x p^x q^{n-x}}{n C_{x-1} p^{x-1} q^{n-(x-1)}}$$

$$= \frac{n! / (x! (n-x)! \cdot p^x q^{n-x})}{n! / ((x-1)! (n-x+1)! \cdot p^{x-1} q^{n-x+1})}$$

$$= \frac{x! (n-x)!}{(x-1)! (n-x+1)!} \cdot \frac{p^x q^{n-x}}{p^{x-1} q^{n-x+1}}$$

$$= \frac{(n-x+1) \cdot p}{x \cdot q}$$

(a) Using the above formula with $n = 8$, $x = 6$ and $P(X = 6)/P(X = 5) = 1/6$ gives

$$3p/6q = 1/6$$

$$3p = q$$

$$3p = 1 - p$$

$$4p = 1$$

$$p = 1/4$$

(b) Let X be the number of damaged tiles, then $X \sim B(12, p)$.

Since $1/3$ of the tiles are seconds, the ratio

seconds : perfect = 1 : 2

$P(X = 1) : P(X = 0) = 1 : 2$

so we use the above formula with $n = 12$, $x = 1$: -

$$12p/q = 1/2$$

$$24p = 1 - p$$

$$25p = 1$$

$$p = 1/25$$

Proportion scrapped = $P(X > 1)$

$$= 1 - P(X = 0 \text{ or } 1)$$

$$= 1 - \{ (24/25)^{12} + 12 \cdot (1/25) \cdot (24/25)^{11} \}$$

$$= 1 - 0.919$$

$$= (\text{say}) \underline{8\% \text{ scrapped.}}$$

Expected profit per table		With probability that $X =$	
£30	£10	0	1
£20	>1		

expected profit from 100 tables

$$= 100[£30 \times (24/25)^{12} + £10 \times 12 \times (1/25) \times (24/25)^{11}$$

$$- £20 \times \{ (1 - P(X=0,1)) \}$$

$$= 100[£30 \times 0.6127 + £10 \times 0.3064 - £20 \times 0.0809]$$

$$= \underline{£1980} \quad 3 \text{ significant figures.}$$

EXAMPLE 4

An employee of a mail order club receives £10 commission for each new customer he obtains. The probability that any prospective client is persuaded to join the club is 0.05.

- Calculate the probability that he will obtain 3 or more members if he knocks on 20 doors.
- State the commission he could expect from knocking on 20 doors.
- State how many doors he would have to knock on per week if he were to obtain 15 new customers per week.
- Calculate the number of doors he must knock on to be 90% sure of getting at least one new customer.

SOLUTION

Let X be the number of new customers, then $X \sim B(20, 1/20)$

$$(a) P(X \geq 3) = 1 - P(X = 0, 1 \text{ or } 2)$$

$$= 1 - 0.925$$

$$= 0.075$$

(b) Expected number of new customers = mean of the Binomial distribution

$$= 20 \times 1/20$$

$$= 1 \text{ new customer.}$$

Therefore expected commission = £10.

(c) Now $X \sim B(n, 1/20)$ and $E(X) = 15$

Therefore $n \times 1/20 = 15$

$$n = 300$$

(d) Again $X \sim B(n, 1/20)$ and we require $P(X \geq 1) \geq 0.9$

$$\text{i.e. } 1 - P(X = 0) \geq 0.9$$

$$1 - (19/20)^N \geq 0.9$$

$$(19/20)^N \leq 0.1$$

$N = 45$. (using trial & error, by calculator or by logs.)

EXERCISES

1. A darts player has a probability of 1/10 of hitting the bull.
Calculate the probability that, with 6 darts he scores
(a) 1, (b) at least 1 bull.

2. In the manufacture of carpets it was found that 4% were below standard. Find the probability that, in a sample of 10: -
(a) 2, (b) not more than 2 were not up to standard.

3. If $X \sim B(n, p)$ and $E(X) = 4$ and $\text{Var}(X) = 2.4$, calculate n and p . - Find $P(X < 4)$

4. The probability that an "A" level maths candidate will pass is 0.7. Calculate the probabilities that, out of 20 candidates:-
(a) all pass
(b) 14 pass
(c) at least 14 pass
(d) not more than 10 fail

5. Show that if $X \sim B(n, p)$ then the ratio $x + 1$ successes to x successes is
$$\frac{(n - x)p}{(x + 1)(1 - p)}$$

Lettuce plants are planted in batches of 12. The number of plants not maturing, X , in 200 such batches is tabulated below.

Fit a Binomial distribution to the data.

Number of plants X	Frequency f
0	73
1	78
2	36
3	10
4	3

6. A multiple choice test consists of 60 questions each with 5 possible answers. If X is the number of correct solutions from a candidate who guesses the answers at random, state the mean and standard deviation for X .
Calculate the probability that such a candidate scores this mean score.
7. Define the Binomial distribution and derive its mean and variance.
8. Playing cards are drawn at random, with replacement from a normal pack of 52.
If the probability that at least one spade is drawn is to be greater than 0.95, how many cards need to be drawn?
9. Fireworks leave the factory in batches of 10,000. If, on average, 0.5% are duds find the mean and standard deviation of the number of duds per batch.
10. A pupil is early for school nine mornings out of ten.
 - (a) Find the expected value and standard deviation of the number of times he is early in a week. (5 days)
 - (b) Find the probability that he will be early at least 4 times in a week.
 - (c) Given that he is early on Monday, find the probability that he will be early at least 4 times that week.

The Poisson Distribution

A discrete random variable X with a probability distribution function (p.d.f.) of the form:

$$P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

is said to be a Poisson random variable with parameter λ . We write $X \sim \text{Po}(\lambda)$

Expectation and Variance

If $X \sim \text{Po}(\lambda)$, then:

- $E(X) = \lambda$.
- $\text{Var}(X) = \lambda$.

Sums of Poissons

Suppose X and Y are independent Poisson random variables with parameters λ and μ respectively. Then $X + Y$ has a Poisson distribution with parameter $\lambda + \mu$. In other words:

- If $X \sim \text{Po}(\lambda)$ and $Y \sim \text{Po}(\mu)$, then $X + Y \sim \text{Po}(\lambda + \mu)$

Random Events

The Poisson distribution is useful because many random events follow it.

If a random event has a mean number of occurrences λ in a given time period, then the number of occurrences within that time period will follow a Poisson distribution. For example, the occurrence of earthquakes could be considered to be a random event. If there are 5 major earthquakes each year, then the number of earthquakes in any given year will have a Poisson distribution with parameter 5.

Example

There are 50 misprints in a book which has 250 pages. Find the probability that page 100 has no misprints.

The average number of misprints on a page is $50/250 = 0.2$. Therefore, if we let X be the random variable denoting the number of misprints on a page, X will follow a Poisson distribution with parameter 0.2.

Since the average number of misprints on a page is 0.2, the parameter, λ of the distribution is equal to 0.2.

$$P(X=0) = (e^{-0.2})(0.2^0)$$

$$= 0.819 \text{ (3sf)}$$

Binomial Approximation

- The Poisson distribution can be used as an approximation to the binomial distribution.
- A Binomial distribution with parameters n and p can be approximated by a Poisson distribution with parameter np .

THE POISSON DISTRIBUTION

DEFINITION

If the random variable X represents the number of occurrences of some "rare event" (i.e. the probability of the event occurring is low) such that

$$P(X = x) = e^{-a} \cdot \frac{a^x}{x!}$$

for $x = 0, 1, 2, 3, \dots, \infty$

then X has a Poisson distribution with mean a .
Or, in short $X \sim Pn(a)$

MEAN AND VARIANCE

$$E(X) = a$$

$$\text{and Var}(X) = a.$$

POISSON APPROXIMATION TO BINOMIAL

If $X \sim B(n, p)$ then, provided n is large (≥ 30 ish) and p is small, (such that $np \leq 6$ ish), then $X \div Pn(a = np)$.

EXAMPLE 1

200 samples, each of 50 screws, are to be drawn from a production line. If 5% of the whole production is defective, calculate the number of samples in which you would expect

- (a) 0, (b) 1, (c) 2, (d) 3 or more defectives

SOLUTION

Let X be the number of defective screws per sample. Then
 $X \sim B(50, 0.05)$. Since n is large and p is small, the Poisson approximation is $X \sim Pn(2.5)$

$$(a) \quad P(X = 0) = e^{-2.5} \cdot \frac{2.5^0}{0!} = e^{-2.5}$$

$$= e^{-2.5}$$

$$= 0.082$$

$$\text{No. of samples} = 200 \times 0.082 = 16$$

$$(b) \quad P(X = 1) = e^{-2.5} \cdot \frac{2.5^1}{1!}$$

$$= 2.5 \times e^{-2.5}$$

$$= 0.205$$

$$\text{No. of samples} = 200 \times 0.205 = 41$$

$$(c) \quad P(X = 2) = e^{-2.5} \cdot \frac{2.5^2}{2!}$$

$$= \frac{2.5}{2} \times e^{-2.5}$$

$$= 0.257$$

$$\text{No. of samples} = 200 \times 0.257 = 51$$

$$(d) \quad P(X \geq 3) = 1 - P(X = 0, 1, 2)$$

$$= 1 - (0.082 + 0.205 + 0.257)$$

$$= 0.456$$

$$\text{No. of samples} = 200 \times 0.456 = 92$$

(Cumulative Poisson probability tables could have been used throughout this question)

EXAMPLE 2

Show that, if $X \sim P(a)$, then $\frac{P(X = x + 1)}{P(X = x)} = \frac{x+1}{a}$

Fast experience has shown that, on average, a motorist has 2.8 punctures per year. Find the probabilities that in any year he has: -

- (a) 0, (b) 1, (c) 2, (d) 3, (e) 4, (f) 5 or more punctures.

If he owns a car for 18 months, state the expected number of punctures and find (g) the probability that he has at least 2 punctures in this period.

SOLUTION

$$\begin{aligned} \frac{P(X = x + 1)}{P(X = x)} &= \frac{e^{-a} \cdot a^{x+1} / (x+1)!}{e^{-a} \cdot a^x / x!} \\ &= \frac{a^{x+1} \cdot x!}{a^x \cdot (x+1)!} \\ &= \frac{a}{x+1} \end{aligned}$$

Let X be the number of punctures per year, then $X \sim Pn(2.8)$

$$\begin{aligned} \text{(a) } P(X = 0) &= e^{-2.8} = 0.0608 \\ \text{(b) } P(X = 1) &= 0.0608 \times 2.8 = 0.1703 \\ \text{(c) } P(X = 2) &= 0.1703 \times 2.8/2 = 0.2384 \\ \text{(d) } P(X = 3) &= 0.2384 \times 2.8/3 = 0.2225 \\ \text{(e) } P(X = 4) &= 0.2225 \times 2.8/4 = 0.1557 \\ \text{(f) } P(X \geq 5) &= 1 - P(X = 0, 1, 2, 3 \text{ or } 4) = 0.1523 \\ \text{(g) } P(Y \geq 2) &= 1 - P(Y = 0 \text{ or } 1) = 1 - e^{-4.2}(1 + 4.2) = 0.922 \end{aligned}$$

If Y is the number of punctures in 1.5 years, then $Y \sim Pn(1.5 \times 2.8)$, $Y \sim Pn(4.2)$

EXAMPLE 3

If the random variables X and Y are independently distributed as Poisson distributions with means a and b respectively, state the distribution of $X + Y$.
Verify this result for $X + Y = 0, 1$ and 2 .

Fred has, on average, a road accident every 4 years.
His wife Betty averages one every 8 years independent of Fred.
Calculate the probabilities that in a given year

- Fred has one accident,
- Betty has one accident,
- They have a total of 2 accidents between them.

SOLUTION

If $X \sim Pn(a)$ and $Y \sim Pn(b)$ then $X + Y \sim Pn(a + b)$

$$P(X + Y = 0) = P(X = 0 \text{ and } Y = 0)$$

$$= e^{-a} \cdot e^{-b}$$

$$= e^{-(a+b)}$$

$$P(X + Y = 1) = P(X = 0 \text{ and } Y = 1 \text{ or } X = 1 \text{ and } Y = 0)$$

$$= e^{-a} \cdot be^{-b} + ae^{-a} \cdot e^{-b}$$

$$= (a + b) \cdot e^{-(a+b)}$$

$$P(X + Y = 2) = P(X = 2 \text{ and } Y = 0 \text{ or } X = 1 \text{ and } Y = 1 \text{ or } X = 0 \text{ and } Y = 2)$$

$$= \frac{a^2}{2} \cdot e^{-a} \cdot e^{-b} + a \cdot e^{-a} \cdot b \cdot e^{-b} + e^{-a} \cdot \frac{b^2}{2} \cdot e^{-b}$$

$$= \frac{e^{-(a+b)}}{2} \cdot (a^2 + 2ab + b^2)$$

$$= \frac{e^{-(a+b)}}{2} \cdot (a + b)^2$$

and each of these probabilities is consistent with $X + Y$ having a Poisson distribution with mean $a + b$.

Let X be the number of accidents Fred has per year, then $X \sim Pn(1/4)$.
If Y is the number of accidents that Betty has per year, $Y \sim Pn(1/8)$.
If $X + Y$ is the total number of accidents they have per year,
 $X + Y \sim Pn(3/8)$.

$$(a) P(X = 1)$$

$$= e^{-0.25} \times 0.25$$

$$= 0.1947$$

$$(b) P(Y = 1)$$

$$= e^{-0.125} \times 0.125$$

$$= 0.1103$$

$$(c) P(X + Y = 2)$$

$$= e^{-0.375} \times 0.375^2/2$$

$$= 0.0483$$

142

EXAMPLE 4

If on average, 2% of rats are albinos, find the probability that there are no albinos in a sample of (a) 10, (b) 100. (c) How large must the sample be to be 99% certain that it contains at least 1 albino.

SOLUTION

(a) Let X be number of albinos in a sample of 10. Then $X \sim B(10, 0.02)$
 $P(X = 0) = 0.98^{10} = 0.817$ (Poisson approximation gives 0.819)

(b) If Y is the number of albinos in a sample of 100, then
 $Y \sim B(100, 0.02)$
 Now since n is large and np is small, the Poisson approximation can be used

$$Y \sim Pn(2) \\ P(Y = 0) = e^{-2} = 0.135$$

(c) If Z is the number of albinos in a sample of n , then, $Z \sim B(n, 0.02)$ and, providing n is large, $Z \sim Pn(0.02n)$

We require $P(Z \geq 1) \geq 0.99$

$$1 - P(Z = 0) \geq 0.99$$

$$1 - e^{-0.02n} \geq 0.99$$

$$e^{-0.02n} \leq 0.01$$

$$e^{0.02n} \geq 100$$

$$n \geq 231$$

(calculator trial and error or logs)

EXERCISES

1. A typist makes, on average, 1.2 errors per page. Find the probabilities that, on a given page he makes
(a) 0, (b) 1, (c) 2, (d) at least one mistake.

2. State how, and under what conditions, the Poisson distribution may be used as an approximation to a Binomial distribution.

A farmer finds that, over a long period of time, one egg in 200 has a double yoke.

Find the probability that in a tray of 30 eggs there are
(a) 0, (b) 1, (c) more than one double yokes.

3. The number of calls arriving at a police station switchboard is one every 30 seconds. Calculate the probability that, in a period of 5 minutes, only 8 calls are received.

4. It is thought that the number of children per family follows a Poisson distribution. The following sample was drawn to test this hypothesis.

Number of children	Number of families
0	12
1	17
2	7
3	4
4	5
5	6
6	1
7	2

Calculate the mean and standard deviation for the number of children per family and thus calculate the theoretical Poisson frequencies.

Plot, on the same axes, frequency polygons to compare the realised and theoretical frequencies.

5. If $X \sim Pn(a)$ show that the ratio
 $P(X = x) : P(X = x - 1) = a : x, (x > 0)$

If $X \sim Pn(3.2)$ obtain successive Poisson probabilities until $P(X = x) < 0.1, (x > 0)$

6. A small supermarket has 2 checkouts. It is found that the first checkout uses, on average 1.2 till rolls per week, the second 1.3 per week.
Calculate the probabilities that 4 or more rolls are used in a particular week by: -
(a) the first checkout,
(b) the second checkout,
(c) the shop.
If both tills run out of rolls at the end of a week, calculate the number of rolls the shopkeeper must have in stock if he is to be at least 95% sure that he does not run out during the following week. (You may assume that the rolls can be inter-changed, if necessary.)
7. Derive the mean and variance of the Poisson distribution.
8. If $X \sim B(n, 0.003)$ and $E(X) = 3$, calculate $P(X > 3)$.
9. Mispprints in the Guardian Newspaper occur on average at a rate of 1 every 2 columns. Calculate the probabilities that a column has
(a) 1, (b) 2, (c) at least 1 misprint.
A page has 8 columns. Calculate the probabilities that, on page 3, there will be
(d) 5, (e) less than 5 misprints.
Saturday's paper contains 24 pages. Calculate the probability that there will be not more than 100 misprints in that edition. (see chapter 7 for Normal approximation)
10. Two golfers A and B find that on average they score 2 and 3 "eagles" respectively per year. Calculate the probabilities that, in a given year,
(a) neither A nor B scores an eagle,
(b) they both score the same number of eagles,
(c) B scores at least 2 eagles more than A.
Give answers correct to 3 decimal places.

POISSON DISTRIBUTION

Exercise A - MEAN + VARIANCE

1. The random variable X has the Poisson distribution with mean 4.8. Find the mean and the variance of (i) $3X - 2$, (ii) $15 - 2X$.
2. The random variable X has the Poisson distribution whose standard deviation is 3. Evaluate, correct to three decimal places,
 - (a) $P(X=9)$, (b) $P(X \geq 9)$, (c) $P(5 \leq X \leq 11)$.
3. The number of times per day that a machine will require attention has the Poisson distribution with mean 1.5. The cost of running the machine on a day when it requires attention X times is $f(12 + 5X^2)$.

- (a) Calculate the expected daily cost of running the machine.
- (b) Find the probability that the cost of running the machine on a randomly chosen day will exceed £22.

Exercise B - POISSON APPROXIMATION TO BINOMIAL

1. Given that $X \sim B(150, 0.04)$, find approximate values for (a) $P(X \leq 3)$, (b) $P(X = 5)$, (c) $P(X > 4)$.
2. The probability that a mass-produced item will be defective is 0.002. The items are packed in boxes, each box containing 500 items. Using an appropriate distributional approximation find, correct to three decimal places, the probability that a box will contain (a) exactly two defective items, (b) at least three defective items.
3. It has been estimated that 0.7% of women have a certain disease. Find, approximately, the probability that in a group of 400 women no more than five have the disease.
4. The probability of a baby being stillborn is 0.0001. Find the probability that of the next 5000 births (a) none will be stillborn, (b) exactly one will be stillborn, (c) at least two will be stillborn.
5. An aeroplane has 196 seats. Past experience has shown that 1.5% of passengers who have booked flights fail to turn up. If the airline sells 200 tickets for a particular flight, find approximate values for the probability that (a) more than 196 turn up, (b) the flight will have one or more empty seats.
6. A certain make of television has a maintenance-free guarantee for four years. The makers are confident that 99.6% of its televisions will operate trouble-free for four years. In a consignment of 200 televisions, find approximate values for the probabilities that (a) none will need any maintenance during the guarantee period, (b) at least 197 of them will not require maintenance during the guarantee period.
7. The probability that a mass produced item will be defective is 0.01. Find approximate values for the probabilities that in a random sample of 200 items
 - (a) exactly 3 will be defective,
 - (b) at most 2 will be defective,
 - (c) at least 3 will be defective.

The number of spectators, X , who pay to watch a village cricket match can be modelled by a Poisson distribution with mean 9.

(a) Write down the standard deviation of X . [1]

(b) Find the probability that, for a randomly chosen match,

(i) exactly 10 spectators come to watch,

(ii) fewer than 10 spectators come to watch. [4]

(c) Each spectator pays a £2 admission charge and the gateman who collects these charges is paid £5. Find the mean and standard deviation of the profit made on admission charges. [5]

(a) Brian and Charles often play table tennis against each other. They know from past experience that Brian wins a game with probability 0.4, independently of all other games. One evening, they play 10 games and X denotes the number of games that Brian wins.

(i) State the distribution of X .

(ii) Determine the mean and variance of X .

(iii) Calculate $P(X = 6)$.

(iv) Determine $P(X \geq 4)$. [8]

(b) The probability that a game of table tennis between Sian and Terri takes less than 10 minutes is 0.06. They play 75 games. Use an approximating distribution to find the probability that more than 5 of these games take less than 10 minutes. [2]

The number of cars, X , entering the car park at a small railway station on a Saturday morning can be modelled by a Poisson distribution with mean 8.

(a) Calculate

(i) $P(X = 6)$,

(ii) $P(4 \leq X \leq 10)$. [5]

(b) The charge per car for parking on a Saturday morning is £4 and the attendant who collects the money is paid £12. The profit from the car park on a Saturday morning is denoted by Y .

(i) Write down an expression for Y in terms of X .

(ii) Find the mean and variance of Y .

(iii) State, with a reason, whether or not Y has a Poisson distribution. [7]

Students in a typing class are each given a page to type. The number of errors made by each student can be modelled by a Poisson distribution with mean μ , where μ varies from student to student.

(a) For Ann, $\mu = 5.1$. Calculate the probability that Ann makes exactly 3 errors. [2]

A page is regarded as unsatisfactory if it contains 4 or more errors.

(b) For Brian, $\mu = 4.5$.

(i) Find the probability that Brian's page is unsatisfactory.

(ii) Given that Brian's page is unsatisfactory, find the probability that it contains 6 errors at most. [7]

(c) The probability that Charlie's page is unsatisfactory is 0.4. Using tables find, to one decimal place, an approximation to the value of μ for Charlie. [2]

(a) When a drawing pin is thrown onto a table, it has probability 0.3 of falling 'point upwards', All 20 drawing pins in a packet are thrown onto a table. Let X denote the number falling 'point upwards'.

- (i) Identify the distribution of X .
- (ii) Find the mean and variance of X .
- (iii) Find $P(5 \leq X \leq 10)$.

(b) For this type of drawing pin, there is a probability of 0.012 that the point is blunt. Large boxes contain 250 of these drawing pins. Find, approximately, the probability that a large box contains at least 3 blunt drawing pins.

(a) A manufacturer produces sweets, 60% of which are red and 40% of which are yellow. A bag contains 10 of these sweets chosen at random. Find the probability that the bag contains

- (i) exactly 3 yellow sweets,
- (ii) more yellow sweets than red sweets.

(b) Another manufacturer produces sweets, 2% of which are soft-centred. A person chooses 60 of these sweets at random. Use the Poisson distribution to find an approximate value for the probability that fewer than 3 of the sweets are soft-centred.

Marian is an archer. When she shoots an arrow at a target, there is a probability of 0.4 that she hits the 'bull'.

(a) Find the probability that she hits the 'bull'.

- (i) exactly 6 times when she shoots 16 arrows,
- (ii) at least 5 times when she shoots 10 arrows.

(b) Marian is invited to a charity event where she shoots 20 arrows at a target. A sponsor agrees to donate £50 to charity each time Marian hits the 'bull'. Calculate the expected value of the donation.

(a) A painter holds a sale of his paintings every Sunday in the town hall. He sells his paintings for £100 each and he pays £50 to hire a room in the town hall. The number of paintings that he sells each Sunday may be assumed to follow a Poisson distribution with mean 1.8.

- (i) Find the probability that the number of paintings that he sells on a particular Sunday lies between 2 and 4 (inclusive).
- (ii) Calculate the mean and variance of the profit made each Sunday from the sale of paintings.

(b) He also takes commissions each Sunday to paint portraits. Assuming that the number of commissions taken follows a Poisson distribution with mean 1.15, calculate (without the use of tables) the probability that he obtains at least 2 commissions on a particular Sunday.

Independently for each seed of a particular type of flower that is sown, the probability that the seed germinates is 0.8.

(a) A gardener plants 10 of these seeds.

- (i) Calculate the mean and standard deviation of the number of these seeds which germinate.
- (ii) Find the probability that at least 6 of these seeds germinate.

(b) Another gardener plants 18 of these seeds. Calculate the probability that exactly 15 of these seeds germinate.

Summary

- For the discrete random variable X with probability density function $P(X=x)$

$$\sum P(X=x) = 1$$

Cumulative distribution function $F(x) = P(X \leq x)$

$$\mu = E(X) = \sum x P(X=x) \quad \text{where } \mu \text{ is the expectation of } X$$

$$E(g(X)) = \sum g(x) P(X=x)$$

$$E(g(x) + h(x)) = E(g(X)) + E(h(X))$$

$$E(X^2) = \sum x^2 P(X=x)$$

$$\sigma^2 = \text{Var}(X) = E(X - \mu)^2 \quad \text{where } \text{Var}(X) \text{ is the variance of } X$$

$$= E(X^2) - \mu^2 \quad (\text{or } \text{Var}(X) = E(X^2) - E^2(X))$$

$$= \sum x^2 P(X=x) - \mu^2$$

σ = standard deviation of $X = \sqrt{\text{Var}(X)}$
- For the random variable X and constants a and b ,

$E(a) = a$	$E(aX + b) = aE(X) + b$
$E(aX) = aE(X)$	$E(aX + b) = aE(X) + b$
$\text{Var}(a) = 0$	$\text{Var}(aX) = a^2 \text{Var}(X)$
$\text{Var}(aX + b) = a^2 \text{Var}(X)$	$\text{Var}(aX + b) = a^2 \text{Var}(X)$
- For any two random variables X and Y and constants a and b ,

$$E(X + Y) = E(X) + E(Y)$$

$$E(X - Y) = E(X) - E(Y)$$

$$E(aX + bY) = aE(X) + bE(Y)$$

$$E(aX - bY) = aE(X) - bE(Y)$$
- For independent random variables X and Y and constants a and b ,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$$

$$\text{Var}(aX - bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$$
- For n independent observations of X ,

$$E(X_1 + X_2 + \dots + X_n) = nE(X)$$

$$\text{Var}(X_1 + X_2 + \dots + X_n) = n \text{Var}(X)$$
- For multiples of X ,

$$E(nX) = nE(X)$$

$$\text{Var}(nX) = n^2 \text{Var}(X)$$

Summary

• Uniform distribution

$$P(X = x_i) = \frac{1}{n} \text{ for } i = 1, 2, \dots, n$$

• Geometric distribution $X \sim \text{Geo}(p)$

p is the probability of a successful outcome

X is the number of independent trials needed to obtain the first successful outcome.

$$P(X = x) = q^{x-1} \times p \text{ for } x = 1, 2, 3, \dots, \text{ to infinity, where } q = 1 - p$$

Note that X cannot be zero.

$$E(X) = \frac{1}{p}, \quad \text{Var}(X) = \frac{q}{p^2}, \quad \text{mode} = 1.$$

• Binomial distribution $X \sim B(n, p)$

p is the probability of a successful outcome.

X is the number of successful outcomes in n independent trials

$$P(X = x) = {}^nC_x q^{n-x} \times p^x \text{ for } x = 0, 1, 2, \dots, n, \text{ where } q = 1 - p$$

$$E(X) = np \text{ and } \text{Var}(X) = npq.$$

• Poisson distribution $X \sim \text{Po}(\lambda)$

X is the number of occurrences of an event in a given interval of time or space, when the mean number of occurrences in the given interval is λ .

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!} \text{ for } x = 0, 1, 2, 3, \dots, \text{ to infinity}$$

$$E(X) = \lambda \text{ and } \text{Var}(X) = \lambda.$$

• Poisson approximation to the binomial distribution

If $X \sim B(n, p)$ with $n > 50$ and $p < 0.1$, then $X \sim \text{Po}(np)$ approximately.

• Sum of independent Poisson variables

If $X \sim \text{Po}(m)$ and $Y \sim \text{Po}(n)$ then $X + Y \sim \text{Po}(m + n)$.

Continuous Random Variables

A continuous random variable is a random variable where the data can take infinitely many values. For example, a random variable measuring the time taken for something to be done is continuous since there are an infinite number of possible times that can be taken.

For any continuous random variable with probability density function $f(x)$, we have that:

$$f(x) \geq 0$$

$$\int_{\text{all } x} f(x) dx = 1$$

This is a very useful result.

$$P(a \leq x \leq b) = \int_a^b f(x) dx$$

Example

X is a continuous random variable with probability density function (p.d.f) given by

$$f(x) = cx \text{ for } 0 \leq x \leq 1,$$

$$= 0 \text{ otherwise.}$$

Where ' c ' is a constant. Find c .

Solution

$$\text{If we integrate } f(x) \text{ between 0 and 1, } \int_0^1 cx dx = \frac{c}{2}$$

Hence $c/2 = 1$ (from the useful result above), giving $c = 2$.

Cumulative Distribution Function (c.d.f.)

If X is a continuous random variable with p.d.f. $f(x)$ defined on $a \leq x \leq b$, then the cumulative distribution function (c.d.f.), written $F(x)$ is given by:

$$F(x) = P(X \leq x) = \int_a^x f(t) dt$$

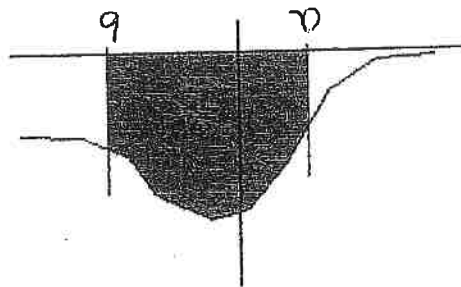
So the c.d.f. is found by integrating the p.d.f. between the minimum value of a and x .

Similarly, the probability density function of a continuous random variable can be obtained by differentiating the cumulative distribution.

The c.d.f. can be used to find out the probability of a random variable being between two values:

$P(a \leq X \leq b)$ = the probability that X is between ' a and b '. But this is equal to the probability that $X \leq b$ minus the probability that $X \leq a$.

[We want the probability that X is in the red area.]



Hence:

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

Expectation and Variance

With discrete random variables, we had that the expectation was $\sum x P(X=x)$, where $P(X=x)$ was the p.d.f.. It may come as no surprise that to find the expectation of a continuous random variable, we integrate rather than sum, i.e.:

$$E(X) = \int x f(x) dx$$

$$E(X^2) = \int x^2 f(x) dx$$

As with discrete random variables, $\text{Var}(X) = E(X^2) - [E(X)]^2$

CONTINUOUS RANDOM VARIABLES

The following are examples of continuous random variables:

- the mass, in grams, of a bag of sugar packaged by a particular machine
- the time taken, in minutes, to perform a task,
- the height, in centimetres, of a five-year-old girl,
- the lifetime, in hours, of a 100-watt light bulb.

PROBABILITY DENSITY FUNCTION (P.D.F.)

A continuous random variable X is given by its probability density function (p.d.f.), which is specified for the range of values for which x is valid. The function can be illustrated by a curve, $y = f(x)$. Note that this function cannot be negative throughout the specified range. Probabilities are given by the area under the curve. It is sometimes possible to find an area by geometry, for example by using formulae for the area of a triangle or a trapezium. Often, however, areas need to be calculated using integration.

Example 1

X is the delay, in hours, of a flight from Chicago, where

$$f(x) = 0.2 - 0.02x, \quad 0 \leq x \leq 10$$

Find

- the probability that the delay will be less than four hours,
- the probability that the delay will be between two and six hours.

Example 2

X is the continuous variable, the mass, in kilograms, of a substance produced per minute in an industrial process, where

$$f(x) = \begin{cases} \frac{1}{32}x(6-x) & (0 \leq x \leq 6) \\ 0 & \text{otherwise} \end{cases}$$

Find the probability that the mass is more than 5 kg.

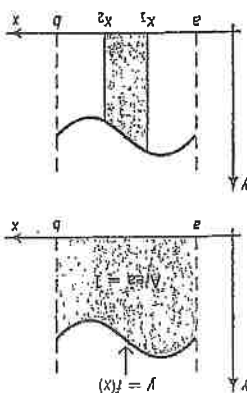
In general, for a continuous random variable X , with p.d.f. $f(x)$

valid over the range $a \leq x \leq b$

$$(a) \int_a^b f(x) dx = 1$$

$$(b) \text{ for } a \leq x_1 \leq x_2 \leq b$$

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x) dx$$



Remember that in an experimental approach, the area under the histogram represents frequency. In a theoretical approach, the area under the curve $y = f(x)$ represents probability.

Example 3.

A continuous random variable has p.d.f. $f(x) = kx^2$ for $0 \leq x \leq 4$.

- Find the value of the constant k .
- Find $P(1 \leq X \leq 3)$.

Example 4.

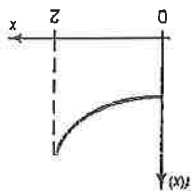
The continuous random variable X has p.d.f. $f(x)$ where

$$f(x) = \begin{cases} k(x+2)^2 & -2 \leq x < 0 \\ 4k & 0 \leq x \leq 1\frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

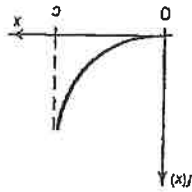
- Find the value of the constant k .
- Sketch $y = f(x)$.
- Find $P(-1 \leq X \leq 1)$.
- Find $P(X > 1)$.

Exercise Calculating probabilities

- The continuous random variable X has a p.d.f. $f(x)$ where $f(x) = kx^2$, $0 \leq x \leq 2$.
 - Find the value of the constant k .
 - Find $P(X \geq 1)$.
 - Find $P(0.5 \leq X \leq 1.5)$.
- The continuous random variable X has p.d.f. $f(x)$ where $f(x) = k$, $-2 \leq x \leq 3$.
 - Sketch $y = f(x)$.
 - Find the value of the constant k .
 - Find $P(-1.6 \leq X \leq 2.1)$.
- The continuous random variable X has p.d.f. $f(x)$ where $f(x) = k(x+2)^2$, $0 \leq x \leq 2$.
 - Find the value of the constant k .
 - Find $P(0 \leq X \leq 1)$ and hence find $P(X > 1)$.



- The continuous random variable X has p.d.f. $f(x)$ where $f(x) = kx$, $0 \leq x \leq c$ and $P(X \leq \frac{1}{2}) = \frac{1}{16}$.
 - Find the values of the constants c and k .
- A continuous random variable has p.d.f. $f(x)$ where $f(x) = kx$, $0 \leq x \leq 4$.
 - Find the value of the constant k .
 - Sketch $y = f(x)$.
 - Find $P(X \leq 1)$.
 - Sketch $y = f(x)$.
 - Find $P(1 \leq X \leq 2.5)$.
 - Find $P(1 \leq X \leq 2.3)$.



Hypothesis Testing

NULL AND ALTERNATIVE HYPOTHESES

Bags of crisps claim to weigh 25g. A sample size of 10 bags gave a sample mean of 23g. Does this show that the crisp manufacturers are ripping off the public?

We need to set up a hypothesis test. A hypothesis test consists of a NULL HYPOTHESIS (H_0) and an ALTERNATIVE HYPOTHESIS (H_1).

The null hypothesis assumes the bag of crisps is 25g and we write:

$$H_0 : \mu = 25g$$

The alternative hypothesis is: $H_1 : \mu < 25g$

We are not always testing a mean – sometimes we need to test a probability or proportion.

Example

Set up hypotheses to test whether a dice is biased when throwing sixes.

$$H_0 : p = \frac{1}{6}$$

$$H_1 : p \neq \frac{1}{6}$$

So we are assuming dice is NOT biased in the null hypothesis.

If H_1 is testing $>$ or $<$ we say the test is ONE SIDED.

If H_1 is testing $\neq$ we say that the test is TWO SIDED.

We then need to test how likely we are to get the value of our sample assuming H_0 is correct. If the probability of getting a value as least as extreme as our sample is very small then we can say it is very unlikely that H_0 is correct and we can reject H_0 in favour of H_1 . To decide whether we can reject H_0 we calculate a p-value.

The p-value

Having set up appropriate null and alternative hypotheses the next step is to formulate a rule to apply to the sample observations to decide whether or not the null hypothesis should be rejected in favour of the alternative hypothesis.

We have to decide on a TEST STATISTIC. This could be:

$$x \sim \text{Bin}(n, p) \quad \text{or} \quad x \sim p_0(\lambda)$$

We then work out the probability of getting a value at least as extreme as our observed sample value (x or $\bar{x}$).

This probability is called the p-value.

Deciding on how small the p-value should be to justify rejecting H_0 depends on the seriousness of the consequences of wrongfully rejecting H_0 . When it is not possible to assess these consequences a rule that is commonly used is to reject H_0 on if the p-value is less than or equal to 0.05.

A more detailed set of guidelines is:-

If $0.01 < \text{p-value} \leq 0.05$ the sample provides STRONG EVIDENCE for rejecting H_0 .
 If $\text{p-value} \leq 0.01$ the sample provides VERY STRONG EVIDENCE for rejecting H_0 .

When $p > 0.05$ all that can be concluded is that there is INSUFFICIENT EVIDENCE TO JUSTIFY REJECTING H_0 . It does not mean H_0 is definitely true.

Working out the p-value

If $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$
 p-value = $p(\bar{x} \geq \bar{x})$ assuming $\mu = \mu_0$

If $H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$
 p-value = $p(\bar{x} \leq \bar{x})$ assuming $\mu = \mu_0$

If $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$
 If $\bar{x} > \mu_0$ work out $p(\bar{x} \geq \bar{x})$ and then multiply by 2
 If $\bar{x} < \mu_0$ work out $p(\bar{x} \leq \bar{x})$ and then multiply by 2

Hypothesis Testing a Probability or Proportion

It is suspected that a coin is biased towards landing head upwards. In 50 tosses of the coin a head occurred 29 times. Test the claim that the coin is biased.

1. Null and alternative hypotheses

$$H_0: p = 0.5$$

$$H_1: p > 0.5$$

p = probability of landing head upwards

2. Distribution of the test statistic

$$x \sim \text{Bin}(50, 0.5)$$

x = no. of heads

3. Observed value

$$x = 29$$

4. Calculation of p-value

$$p(x \geq 29) = 0.1611$$

5. Conclusion

Insufficient evidence to reject H_0 since p value $> 5\%$.
Insufficient evidence to suggest coin is biased in favour of heads.

Hypothesis Testing a Probability or Proportion

Example 1

In the last election only 10% of the Lanelli electorate voted for the Conservatives. Michael Howard is now confident that, with a Lanelli boy as leader of the party, this percentage has increased. During his recent visit a random sample of 20 people were asked how they would vote in the next elections. Four of these said they would vote Conservative. Test whether Michael Howard is justified in his claim that support for the Conservative party has increased in Lanelli.

Example 2

A certain drug is known to cure 55% of patients treated. In a clinical test, 10 patients were given a modified version of the drug which is claimed to be more effective. Given that 8 out of the 10 patients were cured, is one justified in concluding that the modified drug is more effective?

Example 3

a) A supermarket claimed that 25% of its customers paid by credit card. In a random sample of 50 customers it was found that 9 of them paid by credit card. Test the claim made by the supermarket.

b) It was argued that the sample size in (a) was too small and in a further random sample of 160 customers it was found that 28 of them paid by credit card. Once again test the claim made by the supermarket.

Example 1

$$H_0 : p = 0.1$$

$$H_1 : p > 0.1$$

$$x \sim \text{Bin}(20, 0.1)$$

x = no. of people voting conservative

p = probability of someone voting conservative

observed value $x = 4$

$$p(x \geq 4) = 0.1330$$

p -value > 0.05 therefore insufficient evidence to reject the H_0
insufficient evidence to suggest support for conservatives has increased.

Example 2

$$H_0 : p = 0.55$$

$$H_1 : p > 0.55$$

$$x \sim \text{Bin}(10, 0.55)$$

x = no. of patients cured

observed value $x = 8$

$$p(x \geq 8) = p(x = 10) + p(x = 9) + p(x = 8)$$

$$= {}^{10}C_{10}(0.55)^{10}(0.45)^0 + {}^{10}C_9(0.55)^9(0.45)^1 + {}^{10}C_8(0.55)^8$$

$$= 0.0025329 + 0.0207241 + 0.076302$$

$$= 0.09996$$

p value > 0.05 $\therefore$ insufficient evidence to reject H_0 i.e. insufficient evidence to suggest new drug is more effective.

Example 3

a) $H_0 : p = 0.25$

$H_1 : p \neq 0.25$

$x \sim \text{Bin}(50, 0.25)$

$x = \text{no. of people who pay by credit card}$

observed value $x = 9$

$p(x \leq p) = 1 - p(x \geq 10)$

$= 1 - 0.8363$

$= 0.1637$

p value $= 0.3274$

Insufficient evidence to reject H_0

Insufficient evidence to dispute supermarkets claim.

b) $H_0 : p = 0.25$

$H_1 : p \neq 0.25$

$x \sim \text{Bin}(160, 0.25)$

$x = \text{no. of people who pay by credit card}$

observed value $x = 28$

$p(x \leq 28)$

let y be a normal approximation to x

$y \sim N(40, 30)$

$p(y \leq 28.5) = p\left(z < \frac{28.5 - 40}{\sqrt{30}}\right)$

$= p(z < -2.10)$

$= p(z > 2.10)$

$= 0.01786$

p-value $= 0.02572$

p-value < 0.05 therefore strong evidence to reject H_0 and accept H_1 .

Strong evidence to suggest supermarkets claim is incorrect and actually less than 25% use credit card.

Hypothesis Testing a Probability or Proportion

Exercise

1. A lady claimed that she can do better than guessing the suit of a playing card by touch alone. To test her claim, she was blindfolded and asked to identify the suits of 10 playing cards. Test her claim given that she correctly identified the suits of 5 of the cards.
2. It is suspected that a cheap imported die is biased. In 15 throws of the die a 6 occurred once only. Determine whether this information provides sufficient evidence to conclude that the die is biased.
3. The probability that an oyster larva will develop in unpolluted water is 0.9, while in polluted water this probability is less than 0.9. An oyster breeder put 20 larvae in a sample of water and observed that 16 of them developed. Test whether the breeder would be justified in concluding that the water used is polluted.
4. Alan claims that he is a better tennis player than Bill and challenges Bill to a series of games. Test Alan's claim in each of the cases when:
 - a) Alan won 7 of the first 10 games played.
 - b) Alan won 15 of the first 20 games played.

5. In a random sample of 200 adults residing in a certain region it was found that 120 of them were in favour of fluoridation of the domestic water supply. Test the hypothesis that 55% of the adults in the region favour fluoridation against the alternative that the proportion in favour is more than 55%.
6. A leaf disease was reported to affect 25% of all beech trees. A botanist suspects that the incidence of the disease increases with the degree of atmospheric pollution and conducted a survey of beech trees in an urban area. The botanist found 36 infected trees out of 120 examined. Test the botanist's suspicion.
7. A drug company claims in an advertisement that 60% of people suffering from a certain complaint gain instant relief by using their product. A doctor who felt that the claim was rather optimistic gave the product to 200 patients suffering from the complaint and found that 106 of them did gain instant relief. Use this information to test the validity of the advertised claim against the doctor's suspicion that the claim is optimistic.

Past Paper Question

- A darts player claims that he can hit the "bull" in 40% of his throws. A friend is of the opinion that the player's claim is an overestimate and decides to put the player to a test.
- a) In 20 throws, the player hit the "bull" exactly 5 times. Find the p-value of this result and state your conclusion. [3]
 - b) In 150 throws, the player hit the "bull" exactly 48 times. Find an approximation to the p-value of this result and state your conclusion. [5]

Hypothesis Testing 1

The observed value of x was 215, which is less than 240 (the expected value of when $\theta = 0.4$) and H_1 is two-sided.

If the p-value = $2p(x \leq 215 \text{ when } \theta = 0.4)$

Since $n = 600$ is very large we can approximate the distribution $B(600, 0.4)$ by a normal distribution having mean $600 \times 0.4 = 240$ and variance $240 \times 0.6 = 144$. Using this approximation and applying the correction for continuity we have:

$$\text{the p-value} = 2p(x < 215.5 \text{ when } \theta = 0.4) \approx 2p\left(z > \frac{215.5 - 240}{\sqrt{144}}\right)$$

$$\approx 2p(z < -2.04) \approx 2 \times 0.0207 = 0.0414$$

Since the specified significance level of 0.05 is greater than the p-value we reject the null hypothesis and conclude that the proportion of voters supporting Labour has dropped since the last election

Exercise 3.3.2

1. A firm which produces cardigans has found that when a particular machine is operating properly 5% of the cardigans it produces are flawed. An inspection scheme requires 10 cardigans to be selected at random and examined for flaws. The machine is deemed to require adjustment if at least two of the cardigans in the sample are flawed. Find the significance level of this decision rule.
2. The manufacturer of the dog food "Woof" advertised on television that more than 7 dogs out of 10 preferred "Woof" to any other brand of dog food. To test this claim each of 20 dogs was presented with a free choice from a wide selection of brands of dog food including "Woof". The claim is to be accepted as being valid if at least 18 of the dogs opt for "Woof". Find the significance level of this decision rule.
3. A certain drug is known to cure 65% of patients treated. A pharmaceutical company has developed a new drug which is claimed to have a higher cure rate. To test the claim the new drug is to be trialled on 50 patients.
 - a) Set up appropriate null and alternative hypotheses.
 - b) A doctor decided that he would use the new drug if at least 36 of the 50 patients given the new drug were cured. Find the significance level.
 - c) Derive a decision rule to use whose significance level is as close as possible to 0.05 without exceeding it. State the conclusion to be drawn if 40 of the 50 patients given the new drug are cured.
4. A coin is suspected to be such that when tossed a head is more likely than a tail. To test this suspicion the coin is to be tossed 50 times and a conclusion is to be drawn on the basis of the number, X , of heads that are tossed. Construct such a test having significance level as close to 0.05 as is possible. State the actual significance level of your test.

5. A hospital's records showed that 3 out of every 10 casualty patients who reported to the hospital had to wait longer than 20 minutes before being seen by a doctor. The hospital decided to have an additional doctor available in the casualty department, after which it was found that of the next 20 casualties only two had to wait more than 20 minutes before being seen by a doctor. Test whether the extra doctor has decreased the number of casualties having to wait more than 20 minutes, using a significance level of (a) 2% (b) 5%
6. A coin is to be tossed 20 times to test if it is unbiased. Let x denote the number of heads that will be obtained.
 - a) Find the significance level of the decision rule which will reject H_0 if $x \geq 15$ or if $x \leq 5$.
 - b) Given that 6 heads were tossed determine whether it should be concluded that the coin is biased if the significance level is 10%
7. It has been claimed that 10% of the pupils at a large school have part-time jobs. The headmaster is of the opinion that the proportion is greater than 10%. In a random sample of 100 pupils from the school it was found that 15 of them had part-time jobs. Assuming a 5% significance level, is the sample evidence sufficiently strong to support the headmaster's opinion?
8. A gambler claims that he can throw a fair die in such a way that his probability of getting a 6 is greater than $1/6$. To justify his claim he throws the die 50 times. Given that a 6 occurred in 13 of the throws test his claim using a 5% significance level.
9. A manufacturer of colour televisions claims that more than 90% of the sets will be trouble-free for at least two years. In a random sample of 100 such sets that had been sold it was found that 5 of them needed attention during the first two years. Assuming a 5% significance level should the manufacturer's claim be rejected?
10. A large college claims that it admits equal numbers of males and females. A random sample of 400 students at the college was found to consist of 210 males and 190 females. Test the claim using a 5% significance level.
11. In 100 tosses of a coin a head occurred 61 times. Use a 5% significance level to test whether the coin is unbiased.

Exercise 3.3.2

1. $H_0 : p = 0.05$
 $H_1 : p > 0.05$
 $p = \text{probability machine is flawed}$

$x \sim \text{Bin}(10, 0.05)$ $x = \text{machine is flawed}$

observed value $x = 2$

$$p(x \geq 2) = 0.0861$$

significance value = 8.61%

2. $H_0 : p = 0.7$
 $H_1 : p > 0.7$
 $p = \text{probability prefers 'woof'}$

$x \sim \text{Bin}(20, 0.7)$ $x = \text{prefers 'woof'}$

observed value $x = 18$

$$p(x \geq 18) =$$

let $y \sim \text{Bin}(20, 0.3)$

$$p(x \geq 18) = p(y \leq 2)$$

$$= 1 - p(y \geq 3)$$

$$= 1 - 0.9645$$

$$= 0.0355$$

Significance value = 3.55%

1651

3)

a) $H_0: p = 0.65$

$H_1: p > 0.65$

$p = \text{probability of cure}$

$x \sim \text{Bin}(50, 0.65)$ $x = \text{cure}$

b) Observed value $x = 36$

$p(x \geq 36)$

let $y \sim \text{Bin}(50, 0.35)$ $p = \text{probability no cure}$

$p(x \geq 36) = p(y \leq 14)$

$= 1 - p(y \geq 15)$

$= 1 - 0.8122$

$= 0.1878$

Significance level = 18.78%

c) 0.05

$= 1 - 0.95$

$= 1 - p(y \geq 12) = p(y \leq$

$= p(x \geq 39)$

Decision is to accept claim if $x \geq 39$.
If 40 out of 50 were cured you would accept the claim

16/9

4)

$H_0 : p = 0.5$
 $H_1 : p > 0.5$
 $p = \text{probability of heads}$

$x \sim \text{Bin}(50, 0.5)$ $x = \text{head}$

Significance level 0.05

$= p(x \geq 31)$

$= 0.0595$

$= 5.95\% \text{ significant level}$

Accept claim if 31 or more heads are thrown out of the 50.

5)

$H_0 : p = 0.3$
 $H_1 : p < 0.3$
 $p = \text{probability had to wait more than 20 mins}$

$x \sim \text{Bin}(20, 0.3)$ $x = \text{had to wait more than 20 mins}$

observed value $x = 2$

$p(x \leq 2) = 1 - p(x \geq 3)$

$= 1 - 0.9645$

$= 0.0355$

a) At the 2% significance level there is insufficient evidence to reject H_0 and accept H_1 .
 There is insufficient evidence to say he has decreased the number of casualties
 having to wait more than 20 minutes.

b) At the 5% significance level there is strong evidence to reject H_0 and accept H_1 .
 There is strong evidence to suggest he has decreased the number of casualties
 having to wait more than 20 minutes.

167

6)

$$H_0 : p = 0.5$$

$$H_1 : p \neq 0.5$$

p = probability of head

$$x \sim \text{Bin}(20, 0.5)$$

x = no. of heads

$$a) \quad p(x \geq 15) = 0.0207$$

$$p(x \leq 5) = 1 - p(x \geq 6)$$

$$= 1 - 0.9793$$

$$= 0.0207$$

$$\text{Significance level} = 0.0414$$

b) observed value $x = 6$

$$p(x \leq 6) = 1 - p(x \geq 7)$$

$$= 1 - 0.9423$$

$$= 0.0577$$

$$p\text{-value} = 0.1154$$

$p\text{-value} > 0.1$ therefore at the 10% significance level there is insufficient evidence to reject H_0 . Insufficient evidence to suggest coin is biased.

7)

$$H_0: p = 0.1$$

$$H_1: p > 0.1$$

$$x \sim \text{Bin}(100, 0.1) \quad x = \text{has part time job}$$

$$\text{observed value } x = 15$$

$$p(x \geq 15) =$$

let y be a normal approximation to x

$$y \approx N(10, 9)$$

$$p(x \geq 15) = p(y > 14.5)$$

$$= p\left(z > \frac{14.5 - 10}{\sqrt{9}}\right)$$

$$= p(z > 1.5)$$

$$= 0.0668$$

p-value > 0.05 therefore at the 5% significance level there is insufficient evidence to reject H_0 . Insufficient evidence to suggest more than 10% have part time jobs.

8)

$$H_0: p = \frac{1}{6}$$

$$H_1: p > \frac{1}{6}$$

$$p = \text{probability of getting a six}$$

$$x \sim \text{Bin}\left(50, \frac{1}{6}\right) \quad x = \text{no. of sixes}$$

$$\text{observed value } x = 13$$

$$p(x \geq 13)$$

let y be a normal approximation to x

$$y \approx N\left(\frac{25}{3}, \frac{125}{18}\right)$$

$$p(x \geq 13) = p(y > 12.5)$$

$$= p\left(z > \frac{12.5 - \frac{25}{3}}{\sqrt{\frac{125}{18}}}\right)$$

$$= p(z > 1.58)$$

$$= 0.0571$$

p-value > 0.05 therefore at the 5% significance level there is insufficient evidence to reject H_0 . Insufficient evidence to suggest he is more likely to get a six.

168/

9)

$H_0 : p = 0.5$
 $H_1 : p > 0.5$
 $p = \text{probability of male student}$

$X \sim \text{Bin}(400, 0.5)$ $X = \text{no. of male students}$

observed value $x = 210$

$p(x \geq 210)$

let y be a normal approximation to x

$y \approx N(200, 100)$

$p(x \geq 210) = p(y > 209.5)$

$= p(z > \frac{209.5 - 200}{\sqrt{100}})$

$= p(z > 0.95)$

$= 0.1711$

As $p\text{-value} > 0.05$ there is insufficient evidence to reject H_0 . Insufficient evidence to suggest claim is wrong.

11)

$H_0 : p = 0.5$

$p = \text{probability of a head}$

$H_1 : p \neq 0.5$

$x \sim \text{Bin}(100, 0.5)$

observed value $x = 61$

$p(x \geq 61) =$

let y be a normal approximation to x

$y \approx N(50, 25)$

$p(x \geq 61) = p(y > 60.5)$

$= p(z > \frac{60.5 - 50}{\sqrt{25}})$

$= p(z > 2.1)$

$= 0.01786$

$p\text{-value} = 0.03572$

As $p\text{-value} < 0.05$ there is strong evidence to reject H_0 . Strong evidence to suggest coin is biased.

164/

Hypothesis Testing 1

Since the p-value is less than 0.05 but greater than 0.01 there is strong evidence against H_0 so we can conclude that there has been a change in the mean distance travelled. Since the observed $\bar{x}$ is less than $E(x)$ we can be more positive and conclude that the mean distance travelled after the introduction of the new driving regulations is less than the previous mean distance travelled.

Exercise 3.2.1

1. A random sample of 25 observations is taken from a normal distribution having mean μ and standard deviation 2 in order to test $H_0: \mu = 10$ against $H_1: \mu > 10$. Carry out the test given that the observed sample mean was 11.

2. When a machine is operating correctly it produces steel pins whose lengths are normally distributed with mean 2.5cm and standard deviation 0.05cm. Any malfunction of the machine affects only the mean length. Samples are taken periodically in order to check that the machine is operating correctly. Each such sample consists of 40 pins. Carry out a test to determine if the machine is malfunctioning given that the sample mean was (a) 2.49cm (b) 2.52cm.

3. A company manufacturing electric light bulbs has a research section for developing modifications to the manufacturing process aimed at increasing the mean lifetime of the bulbs. At present the lifetimes are normally distributed with mean 2200 hours and standard deviation 118 hours. It may be assumed that any modification to the present process will affect the mean lifetime only. Any modification is used to produce 100 bulbs. Test whether a modification should be adopted given that the mean lifetime of the 100 bulbs was (a) 2210 hours, (b) 2231 hours.

4. Items produced by a certain machine have masses that are normally distributed with mean μ kg and standard deviation 0.02kg. If $\mu \neq 1$ then the machine setting should be corrected. Given that the mean mass of a random sample of 25 items was 0.989kg determine whether consideration should be given to correcting the machine-setting.

5. The tar yields in cigarettes of a particular brand are known to be normally distributed with mean μ mg and standard deviation 0.8mg. In a random sample of 10 cigarettes the tar yields were found to be:
17.1, 19.2, 18.3, 18.2, 18.9, 17.8, 18.3, 16.9, 18.5
Test the hypothesis that $\mu = 17.5$ against the alternative that $\mu > 17.5$

Exercise 3.2.1

1. $H_0: \mu = 10$

$H_1: \mu > 10$

$x \sim N(10, 2^2)$

$\bar{x} \sim N(10, \frac{2^2}{25})$

Observed value $\bar{x} = 11$

$p(z \geq 11) = p(z > \frac{11-10}{\sqrt{\frac{2^2}{25}}})$

$= p(z > 2.5)$

$= 0.00621$

$\therefore$ p-value = 0.00621

Since the p-value is smaller than 0.01 there is very strong evidence to reject H_0 and accept H_1 . There is very strong evidence to suggest the mean is greater than 10.

2. $H_0: \mu = 2.5$

$H_1: \mu \neq 2.5$

$x \sim N(2.5, 0.05^2)$

x = length of pins

$\bar{x} \sim N(2.5, \frac{0.05^2}{40})$

a) Observed value $\bar{x} = 2.49$

$p(\bar{x} \leq 2.49) = p(z < \frac{2.49-2.5}{\sqrt{\frac{0.05^2}{40}}})$

$= p(z < -1.26)$

$= p(z > 1.26)$

$= 0.1038$

$\therefore$ p-value = 0.2076

Since the p-value is greater than 0.05 there is insufficient evidence to reject H_0 . There is insufficient evidence to suggest that the mean length of pins has changed/malfunctioned.

b) Observed value $\bar{x} = 2.52$

$$p(\bar{x} \geq 2.52) = p\left(z > \frac{2.52 - 2.5}{\frac{0.05}{\sqrt{40}}}\right)$$

$$= p(z > 2.53)$$

$$= 0.00570$$

$$\therefore \text{p-value} = 0.0114$$

Since the p-value is less than 0.05 there is strong evidence to reject H_0 and accept H_1 . There is strong evidence to suggest that the mean length of pins has changed therefore machine is malfunctioning.

$$3. H_0: \mu = 2200$$

$$H_1: \mu > 2200$$

$$x \sim N(2200, 118^2)$$

x = lifetime of bulbs

$$\bar{x} \sim N\left(2200, \frac{118^2}{100}\right)$$

a) Observed value $\bar{x} = 2210$

$$p(\bar{x} \geq 2210) = p\left(z > \frac{2210 - 2200}{\frac{118}{\sqrt{100}}}\right)$$

$$= p(z > 0.85)$$

$$= 0.1977$$

$$\therefore \text{p-value} = 0.1977$$

Since the p-value is greater than 0.05 there is insufficient evidence to reject H_0 . There is insufficient evidence to suggest the bulbs have a longer lifetime.

$$b) \text{ Observed value } \bar{x} \geq 2231 = p\left(z > \frac{2231 - 2200}{\frac{118}{\sqrt{100}}}\right)$$

$$= p(z > 2.63)$$

$$= 0.00427$$

$$\therefore \text{p-value} = 0.00427$$

Since the p-value is less than 0.01 there is very strong evidence to reject H_0 and accept H_1 . There is very strong evidence to suggest the bulbs have a longer lifetime.

17/2

$$4. H_0: \mu = 1$$

$$H_1: \mu \neq 1$$

$$x \sim N(1, 0.02^2)$$

$$\bar{x} \sim N\left(1, \frac{0.02^2}{25}\right)$$

$$\text{Observed value } \bar{x} = 0.989$$

$$p(x \leq 0.989) = p\left(z < \frac{0.989 - 1}{\sqrt{\frac{0.02^2}{25}}}\right)$$

$$= p(z < -2.75)$$

$$= p(z > 2.75)$$

$$= 0.00298$$

$$\therefore p\text{-value} = 0.00596$$

Since the p-value is less than 0.01 there is very strong evidence to reject H_0 and accept H_1 . There is very strong evidence to suggest the machine-settings should be corrected.

$$5. H_0: \mu = 17.5$$

$$H_1: \mu > 17.5$$

$$x \sim N(17.5, 0.8^2)$$

$$x = \text{tar yield in cigarette}$$

$$\bar{x} \sim N\left(17.5, \frac{0.8^2}{10}\right)$$

$$\text{Observed value } \bar{x} = 18.1$$

$$p(x \geq 18.1) = p\left(z > \frac{18.1 - 17.5}{\sqrt{\frac{0.8^2}{10}}}\right)$$

$$= p(z > 2.37)$$

$$= 0.00889$$

Since the p-value is less than 0.01 there is very strong evidence to reject H_0 and accept H_1 . There is very strong evidence to suggest the tar yield in this brand of cigarettes is greater than 17.5mg.

173/

Hypothesis Testing the Mean of a Poisson Distribution

Example 1

Daily records over a long period of time showed that the number of vehicles requiring assistance per day along a certain motorway has a Poisson distribution with mean 0.5. It was suggested that the mean would be greater on wet days. Observation on 10 wet days showed that a total of 9 vehicles required assistance. Assuming that a Poisson distribution is an appropriate model for the number of vehicles requiring assistance on a wet day, test whether the sample evidence is strong enough to validate the suggestion.

Example 2

In the situation described in example 1 it was found that a total of 40 vehicles required assistance over 55 wet days. Test whether this result is sufficient to justify concluding that the average number of vehicles requiring assistance per wet day is greater than 0.5. Let y denote the total number of vehicles that require assistance on 55 wet days.
Then $y \sim \text{Po}(55, \mu)$ where μ is the average number of vehicles requiring assistance per wet day.

Exercise

1. In the past, the number of claims that were made per week to an insurance company had a Poisson distribution with mean 1.25. In a recent four week period a total of 10 claims were received by the company. Is the company justified in concluding that the mean number of claims per week has increased?
2. A survey of tabloid newspapers showed that the number of misspellings per page had a Poisson distribution with mean 1.4. Five pages of a particular tabloid newspaper contained 3 misspellings. Test whether the mean number of misspellings in this particular tabloid newspaper is less than 1.4.
3. Past records have indicated that the number of accidents along a particular roadway has a Poisson distribution, the average number per month being 10. During a month when additional warning signs had been erected along the roadway the number of accidents that occurred was 5. The Road Safety Officer decided to test whether the signs had been effective in reducing the number of accidents. Having set up appropriate null and alternative hypotheses for the test the officer calculated the p-value to be approximately 0.07. Show how this p-value was obtained and state what the officer's conclusion should be.
4. Records over several years showed that the number of car thefts per week in a small town had a Poisson distribution with mean 1.6. In one particular week the town's local newspaper published an article bearing the heading 'Car thefts on the increase'. The article stated that during the past five weeks there had been a total of 14 car thefts in the town. Test the strength of the evidence for the claim made in the heading of the article.
5. Records of absenteeism at a factory showed that the number of absentees per working day had a Poisson distribution with mean 1.2. During a subsequent week of five working days the number of absentees were 2, 1, 3, 3, 2 respectively. Test whether the evidence of that week is strong enough to conclude that the mean number of absentees per week has changed.

1251

6. The nature of a radioactive source can be identified from the mean rate of emissions. For a type-A source the mean rate of emissions is known to be 2.4 per minute. Given that there were 3 emissions from an unidentified source over a period of four minutes test whether this evidence is sufficiently strong to justify identifying the source as not being of type-A.
7. The number of radios sold per week by a certain shop has a Poisson distribution with mean 5. In the week following the appointment of an additional sales assistant a total of 7 radios were sold. Test whether the additional appointment has increased the mean number of radios sold per week.
8. Experiments were conducted to determine whether an antibiotic spray is effective in reducing the number of bacterial colonies that will develop in dishes of nutrient exposed to an infected environment. It is known that the number of colonies that will develop per unsprayed dish has a Poisson distribution with mean 7.5.
 - a) In an experiment in which one dish of the nutrient was sprayed, the number of colonies that developed was 3. Determine whether this result provides sufficient evidence to conclude that the spray is effective in reducing the mean number of bacterial colonies that develop.
 - b) In another experiment, four dishes were sprayed and the numbers of colonies that developed were 3, 5, 6 and 4 respectively. Determine whether these results provide sufficient evidence to justify concluding that the spray is effective.

Significance Level

When testing the null hypothesis H_0 , the p-value was a measure of how likely it was that H_0 was true. The smaller the p-value the less likely it is that H_0 is true and the stronger the evidence against H_0 .

It is up to the investigator to decide whether a calculated p-value is acceptably small enough to justify the rejection of H_0 . The decision would have to be based on the seriousness of the consequences should H_0 be wrongfully rejected. Ideally, the investigator should be aware of these consequences before the sample is taken and therefore be able to give a value for the probability that he is willing to accept of wrongfully rejecting H_0 .

A pre-specified value for this probability is called the SIGNIFICANCE LEVEL and is usually denoted by α and quoted as a percentage.

In our examples so far the significance level has been 5% since we have rejected H_0 of the p-value was less than or equal to 0.05.

This means there is a 5% chance of wrongfully rejecting H_0 . If the significance level was 10% we would reject H_0 if the p-value was less than or equal to 0.1.

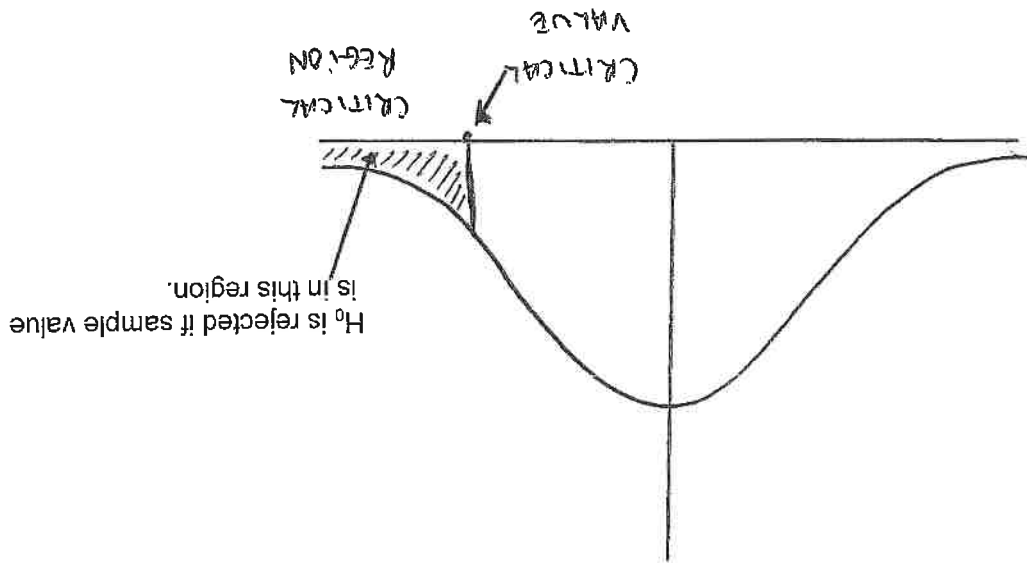
So we reject H_0 if the p-value is $\leq \alpha$ (significance level).

The set of values that lead to the rejection of H_0 is called the critical region.

e.g. If H_0 is $\mu = 25g$

H_1 is $\mu > 25g$

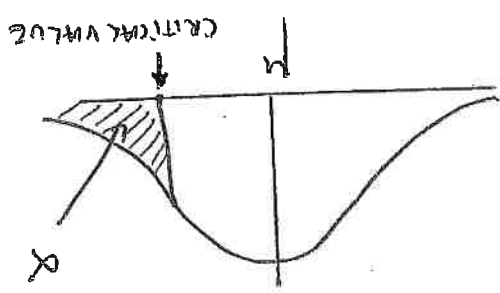
And significance level is 5%, then the critical region such that $p(\bar{x} \geq \bar{x}) = 0.05$



421

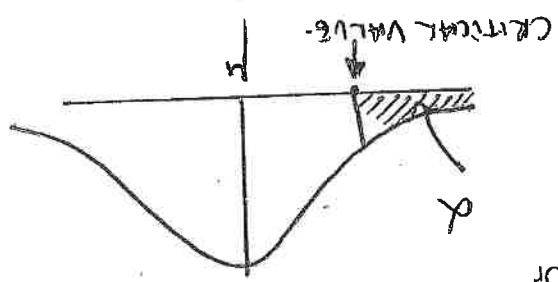
Decision Rule
 Prior to taking a sample we can devise a decision rule for a certain significance level (α).
 This decision rule will give the CRITICAL VALUE and the probability of rejecting H_0 when H_0 is true will be α .

1 sided

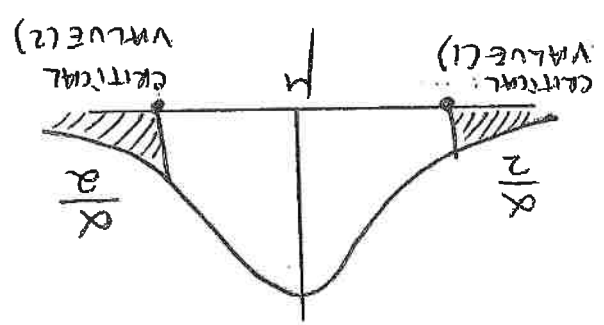


Decision Rule
 Reject H_0 if sample mean
 is $\geq$ CRITICAL VALUE

Or



Decision Rule
 Reject H_0 if sample mean
 is $\leq$ CRITICAL VALUE



Decision Rule
 Reject H_0 if sample mean
 is $\leq$ CRITICAL VALUE (1)
 or $\geq$ CRITICAL VALUE (2)

Significance Testing a Probability

Example 1

The manufacturer of a new type of seeding compost claims that seeds sown in this new compost will have a higher germination rate than if they were sown in ordinary compost. For a particular variety of seed sown in ordinary compost the germination rate is known to be 0.7. To test the manufacturer's claim an experiment was conducted in which 20 seeds were sown in the new compost and the number x that germinated was counted.

- Find the significance level of the decision rule that accepts the claim if $x \geq 17$.
- Find a decision rule with a significance level of 0.05.

Example 2

A coin is to be tossed 50 times to test whether it is unbiased.

- Find the significance level if the decision rule is, conclude that coin is biased if $x \leq 19$ or $x \geq 31$, where x is number of heads obtained.
- Using a 5% significance level what conclusion should be drawn if 32 heads are obtained.

Example 3

In a certain constituency 40% of the electorate voted Labour in the last general election. An opinion poll is to be conducted to spot whether or not there has been a change in the level of support for the Labour party since the last election. In the poll of 600, 215 intended to vote Labour. What conclusion should be drawn if the significance level is 5%?

Significance Testing a Probability

Example 1

$$H_0 : p = 0.7$$

$$H_1 : p > 0.7$$

p = probability seed germinates

$$a) \quad x \sim \text{Bin}(20, 0.7) \quad x = \text{seed germinates}$$

observed value $x = 17$

$$p(x \geq 17)$$

let y = seeds that don't germinate

$$y \sim \text{Bin}(20, 0.3)$$

$$p(x \geq 17) = p(y \leq 3)$$

$$= 1 - p(y \geq 4)$$

$$= 1 - 0.8929$$

$$= 0.1071$$

Significance level = 10.71%

b) 0.05

= 1 - 0.95 Look up 0.95 in table and choose nearest highest value = $p(y \geq 3)$

$$p(y \leq 2) = 1 - p(y \geq 3)$$

$$= 1 - 0.9645$$

$$= 0.0355$$

Decision is to accept claim if $x \geq 18$

17/

Example 2

$$H_0 : p = 0.5$$

$$H_1 : p \neq 0.5$$

$$x \sim \text{Bin}(50, 0.5)$$

$$\text{a) } p(x \leq 19) = 1 - p(x \geq 20)$$

$$= 1 - 0.9405$$

$$= 0.0595$$

$$p(x \geq 31) = 0.0595$$

$$\text{Significance value} = 0.0595 + 0.0595$$

$$= 0.119$$

$$= 11.9\%$$

$$\text{b) observed value } x = 32$$

$$p(x \geq 32) = 0.0325$$

$$p\text{-value} = 2 \times 0.0325$$

$$= 0.065$$

At the 5% significance level there is insufficient evidence to reject H_0 i.e. insufficient evidence to suggest coin is biased.

Example 3

$$H_0 : p = 0.4$$

$$H_1 : p \neq 0.4$$

$$x \sim \text{Bin}(600, 0.4) \quad x = \text{voted labour}$$

observed value $x = 215$

$$p(x \leq 215)$$

let y be a normal approximation to x

$$y \sim N(240, 144)$$

$$p(y \leq 215.5) = p\left(z \leq \frac{215.5 - 240}{\sqrt{144}}\right)$$

$$= p(z \leq -2.04)$$

$$= p(z > 2.04)$$

$$= 0.02068$$

$$\text{p-value} = 0.04136$$

At the 5% significance level there is strong evidence to reject H_0 . There is strong evidence to suggest there has been a change in the level of support for the labour party.

Significance Testing a Poisson Mean

Example 1

The number of parking tickets issued per weekday by the traffic wardens in a town has a Poisson distribution with mean 1.4. The town council decide to employ one more warden on a temporary basis to see if the number of tickets issued will be increased on average. Let x denote the number of tickets issued during the first five weekdays after the employment of the additional warden.

- a) The council decided to conclude that the average number of tickets issued per week had increased if the observed value of x was 10 or more. Find the significance level of the decision rule.
- b) Given that the observed value of x was 11 find the p-value and deduce the minimum significance level for the conclusion to be that the mean number of tickets issued per weekday had increased.

Example 2

Manufactured cloth is passed for sale if the average number of flaws per metre of the cloth is less than 0.5. A decision on whether or not to pass a roll of the cloth is based on the observed number of flaws in a length of 100m from the roll. Denoting this number by x , construct a decision rule having a significance level of approximately 0.01. It may be assumed that the number of flaws has a Poisson distribution.

Example 3

The random variable x has the Poisson distribution with unknown mean μ . A random sample of 3 observations of x is to be taken in order to test $H_0: \mu = 1.7$ against $H_1: \mu < 1.7$.

- a) It is decided to reject H_0 if the sum of the 3 values obtained is less than 3.
- i. Find significance level of this decision rule.
- ii. Find probability that this decision rule will not reject H_0 when true value of μ is 1.2
- b) Find decision rule having a significance level as close as possible to 0.05

Significance Testing a Poisson Mean

Example 1

$$H_0 : \lambda = 1.4$$

$$H_1 : \lambda > 1.4$$

$$a) \quad x \sim \text{po}(7)$$

$x = \text{no. tickets in 5 weekday}$

observed value $x = 10$

$$p(x \geq 10) = 0.1695$$

significance level = 16.95%

b) observed value $x = 11$

$$p(x \geq 11) = 0.0985$$

minimum significance level = 9.85%

Example 2

$$H_0 : \lambda = 0.5$$

$$H_1 : \lambda < 0.5$$

$\lambda = \text{mean no. of flaws per m}$

$$x \sim \text{po}(50) \quad x = \text{no. flaws per 100m}$$

let y be a normal approximation to x

$$\text{let } y \approx N(50, 50)$$

$$Z = \frac{y - \mu}{\sigma} = \frac{x + 0.5 - 50}{\sqrt{50}} = -2.3263$$

decision rule – reject H_0 and accept H_1 if $x \leq 33$

183/

Example 3

a) $H_0: \mu = 1.7$

$H_1: \mu < 1.7$

i. $x \sim \text{po}(\mu)$

$y \sim \text{po}(5.1)$

$y = \text{sum of 3 observations of } x$

observed value = 3

$p(u < 3) = p(y = 0) + p(y = 1) + p(y = 2)$

$$= \frac{e^{-5.1} x^{5.1}}{0!} + \frac{e^{-5.1} x^{5.1}}{1!} + \frac{e^{-5.1} x^{5.1}}{2!}$$

= 0.1165

Significance level = 11.65%

ii. $\mu = 1.2$

$y \sim \text{po}(3.6)$

$p(y < 3) = 1 - p(y \geq 3)$

= $1 - 0.6973$

= 0.3027

$p(\text{not reject } H_0) = 0.6973$

b) $p(y < 2) = p(y = 0) + p(y = 1)$

$$= \frac{e^{-5.3} x^{5.3}}{0!} + \frac{e^{-5.3} x^{5.3}}{1!}$$

Decision rule = reject H_0 if sum of values is < 2 .

This has significance level of 3.72%

184/

2x.

Exercises

- The number of flaws in a given length of cloth has a Poisson distribution. The manufacturer claims that the mean number of flaws per metre is 0.2. To test this claim against the alternative that the mean number of flaws per metre is greater than 0.2, it is decided to examine a length 10cm of the cloth and to reject the claim if it contains 6 or more flaws. Find the significance level of this test. Also find the number that should replace the 6 above so that the new rule will have a significance level of approximately 0.05.
- The number of times that a computer needs attention has a Poisson distribution, the mean number per month being 4.25. During the first month after the computer had undergone a major overhaul there was only one occasion when it needed attention. Using a 5% significance level determine whether it is justifiable to conclude that the major overhaul has reduced the mean number of times that the computer needs attention.
- The number of emergency calls per hour to a gas board over a weekend has a Poisson distribution with a mean 0.7. It was suggested that the mean would be greater than 0.7 over a very cold weekend. During a very cold weekend in the winter the office received 11 emergency calls during the first 10 hours. Show that, with a significance level of 5% this does not provide sufficient evidence to justify concluding that the mean number of calls over that weekend is greater than 0.7. Write down the smallest significance level that would justify the conclusion.
- An employer is seeking a typist who can be relied upon not to mistype more than one character per page on average. An applicant for the position was given four pages to type and it was found that she had mistyped a total of seven characters. Use a 10% significance level to test the hypothesis that her average number of mistyped characters per page is 1 against the alternative that it is greater than 1. Assume that the number of mistyped characters per page has a Poisson distribution.
- It is required to test $H_0: \mu = 2$ against $H_1: \mu \neq 2$, where μ is the mean of a Poisson distribution. The test is to be based on a random sample of five observations from the distribution.
 - Find the significance level of the decision rule which rejects H_0 if the sample mean is < 0.4 or is > 3.4
 - Assuming that the chosen significance level is 0.06, what conclusion should be drawn if the sample mean was 0.8?
- The random variable x has the Poisson distribution with mean μ . A random sample of 23 observations had a sum equal to 30. Use a 5% significance level to test $H_0: \mu = 1.7$ against $H_1: \mu < 1.7$.
- The counts of bacterial colonies that developed in 10 exposed Petri dishes were: 12, 13, 8, 17, 16, 8, 18, 15, 9, 11
 Test, at the 1% significance level, that the mean number of bacterial colonies that will develop per Petri dish under the same exposure is equal to 10 against the alternative that it is greater than 10.

8. It has been claimed that the mean number of impurities per sheet of glass from a certain manufacturer is 2.5. Test this claim, using a significance level of 5%, given that the total number of impurities in 100 sheets of glass was 278. Assume that the number of impurities per sheet of glass has a Poisson distribution.
9. The number of cars arriving at a road junction in 45 consecutive five minute periods were as follows:
- | Number of cars arriving | 0 | 1 | 2 | 3 | 4 | 5 |
|-------------------------|---|----|----|---|---|---|
| Number of time periods | 6 | 13 | 11 | 9 | 3 | 3 |
- Assuming that the number of cars arriving in a five minute period has a Poisson distribution, test, at the 1% level of significance, the hypothesis that the mean of the distribution is equal to 1.5 against the alternative that it is not equal to 1.5.
10. Over a long period of time the number of occasions that a piece of equipment has had to be repaired by a technician has averaged 1.9 times per month. A new technician was employed to be solely responsible for the maintenance of this equipment. It may be assumed that the number of times per month that this technician will be required to repair the equipment has a Poisson distribution. In each of the following cases use a 5% significance level to test whether the mean number of repairs per week is different from 1.9.
- a) In the new technician's first three months he was required to repair the equipment twice.
- b) In the new technician's first thirty months he was required to repair the equipment a total of 43 times.

Exercise

1)

$H_0 : \lambda = 0.2$ λ = mean no. flaws per m cloth

$H_1 : \lambda > 0.2$

$x \sim \text{po} (2)$ x = no. flaws per 10m cloth

observed value $x = 6$

$p(x \geq 6) = 0.0166$

$\therefore$ significance level = 1.66%

$p(x \geq 5) = 0.0527$

decision rule – reject H_0 if ≥ 5

3)

$H_0 : \lambda = 0.7$ λ = mean no. calls per hour

$H_1 : \lambda > 0.7$

$x \sim \text{po} (7)$ x = no. calls per 10 hours

observed value $x = 11$

$p(x \geq 11) = 0.0985$

As $p\text{-value} > 0.05$ there is insufficient evidence at the 5% significance level to reject H_0 . Insufficient evidence to suggest the mean no. calls in that weekend is > 0.7 .

The smallest significance level that would justify the conclusion is $0.0985 = 9.85\%$.

4)

$$H_0: \lambda = 1$$

λ = mean no. mistakes per page

$$H_1: \lambda > 1$$

$$x \sim \text{po}(4)$$

x = no. mistypes per 4 pages

$$p(x \geq 7) = 0.1107$$

observed value $x = 7$

As p-value > 0.1 insufficient evidence to reject H_0

insufficient evidence to suggest the mean no. of mistypes is greater than 1.

5)

$$H_0: \mu = 2$$

$$H_1: \mu \neq 2$$

$$x \sim \text{po}(\mu)$$

$$y \sim \text{po}(10) \quad y = \text{sum of 5 observations}$$

$$\text{a) } p(y < 2) = 1 - p(y \geq 2)$$

$$= 1 - 0.9995$$

$$= 0.0005$$

$$p(y > 17) = p(y \geq 18)$$

$$= 0.0143$$

$$\text{Significance level} = 0.0148 = 1.48\%$$

$$\text{b) observed value } y = 4$$

$$p(y \leq 4) = 1 - p(y \geq 5)$$

$$= 1 - 0.09704$$

$$= 0.0296$$

$$\text{p-value} = 0.0592$$

Using the 0.06 significance level there is strong evidence to reject H_0 .
Strong evidence to suggest $\mu \neq 2$

7)

$$H_0: \lambda = 10 \quad \lambda = \text{mean no. bacterial colonies per dish}$$

$$H_1: \lambda > 10$$

$$x \sim \text{po}(10) \quad x = \text{no. bacterial colonies per dish}$$

$$y \sim \text{po}(100) \quad y = \text{no. bacterial colonies per 10 dishes}$$

$$\text{observed value } y = 127$$

$$p(y \geq 1270)$$

let w be a normal approximation to y

$$w \approx N(100, 100)$$

$$p(y \geq 127) = p(z > \frac{126.5 - 100}{10})$$

$$= p(z > 2.65)$$

$$= 0.00402$$

At the 1% significance level there is strong evidence to suggest the mean no. bacterial colonies per dish is > 10 . Strong evidence to reject H_0 and accept H_1 .

8)

$$H_0: \lambda = 2.5 \quad \lambda = \text{mean no. impurities per sheet of glass}$$

$$H_1: \lambda \neq 2.5$$

$$x \sim \text{po}(2.5) \quad x = \text{no. impurities per sheet}$$

$$y \sim \text{po}(250) \quad y = \text{no. impurities per 100 sheets}$$

$$\text{observed value } y = 278$$

$$p(y \geq 278)$$

let w be a normal approximation to y

$$w \approx N(250, 250)$$

$$p(y \geq 278) = p(w > 277.5)$$

$$= p(z > \frac{277.5 - 250}{\sqrt{250}})$$

$$= p(z > 1.74)$$

$$= 0.0409$$

$$p\text{-value} = 0.0818$$

Using a significance level of 5% there is insufficient evidence to reject H_0 and accept H_1 . Insufficient evidence to suggest the mean no. of impurities per sheet of glass is not 2.5.

1891

9)

$H_0: \lambda = 1.5$ $\lambda = \text{mean no. cars arriving}$

$H_1: \lambda \neq 1.5$

$x \sim \text{po}(1.5)$ $x = \text{no. cars arriving in a 5 min period}$

$y \sim \text{po}(67.5)$ $y = \text{no. cars in 45 consecutive 5 min periods}$

observed value $y = 89$

$p(y \geq 89)$

let w be a normal approximation to y

$w \approx N(67.5, 67.5)$

$p(y \geq 89) = p(w > 88.5)$

$= p(z > \frac{88.5 - 67.5}{\sqrt{67.5}})$

$= p(z > 2.56)$

$= 0.00523$

$p\text{-value} = 0.01046$

using a 1% significance level there is insufficient evidence to reject H_0 . Insufficient evidence to suggest the mean no. cars arriving per 5 minute period is not 1.5.

1901

Type I and Type II Errors

When undertaking a significance test four possible conclusions can be reached. These are summarised as:-

OUR CONCLUSION		TRUE SITUATION	
1	ACCEPT H_0	H_0 IS TRUE	
	WRONG DECISION	H_0 IS TRUE	TYPE I ERROR.
2	REJECT H_0	H_0 IS TRUE	
	WRONG DECISION	H_0 IS TRUE	TYPE II ERROR.
3	ACCEPT H_0	H_0 IS FALSE	
	WRONG DECISION	H_0 IS FALSE	TYPE II ERROR
4	REJECT H_0	H_0 IS FALSE	
	CORRECT DECISION	H_0 IS FALSE	

A Type I error is made if we reject H_0 when H_0 is in fact TRUE.

A Type II error is made if we accept H_0

when H_0 is in fact FALSE.

The associated probabilities of these errors is:-
 $P(\text{Type I error}) = P(\text{Rejecting } H_0 \mid H_0 \text{ is TRUE})$
 $P(\text{Type II error}) = P(\text{accepting } H_0 \mid H_1 \text{ is TRUE})$

Example:
A box is known to contain either
 H_0 : 10 white counters and 90 black counters or
 H_1 : 50 white counters and 50 black counters.
To test the hypothesis H_0 against H_1 four
counters are drawn out at random from the
box, without replacement.

If all 4 counters drawn are black H_0 is accepted
otherwise it is rejected.
Calculate the size of Type I and Type II
errors for this test.

Solution:
 H_0 : The box has 10 white and 90 black counters
 H_1 : " " " " " 50 white " 50 black "

Then:-

$$P(\text{Type I error}) = P(\text{Rejecting } H_0 \mid H_0 \text{ is true}) = P(\text{At least one white} \mid \text{There are 10 white and 90 black in the box})$$

$$= 1 - P(\text{Drawing 4 black counters}) = 1 - \left[\frac{90}{100} \times \frac{89}{99} \times \frac{88}{98} \times \frac{87}{97} \right]$$

$$= 1 - 0.652 = 0.348 = P(\text{Type I error})$$

192

123/

$$P(\text{Type II error}) = 0.059$$

$$P(\text{Drawing 4 black counters}) = \frac{50}{100} \times \frac{49}{99} \times \frac{48}{98} \times \frac{47}{97} = 0.059$$

So black counters

So if the box does contain 50 white and

$$= P(\text{All 4 are black} \mid \text{50 white and 50 black counters in the box})$$

$$P(\text{Type II error}) = P(\text{Accepting } H_0 \mid H_1 \text{ is True})$$

1944

Example 2:
A man claims he can throw a 'six' with a fair cubical die, five times out of six on average.
Find the probability that he will throw four or more sixes in six throws:
(i) if his claim is justified
(ii) if he can throw a six on average,
only once in six throws.
To test the claim, the man is invited to throw the die six times, his claim being accepted if he throws at least four sixes.
Find the probability that the test will
(a) accept the man's claim when the hypothesis (i) is true or (b) reject the claim when it is justified that is when hypothesis (i) is true.

Solution:

(i) If the man's claim is justified
 $P(\text{a six is thrown}) = \frac{5}{6}$

$\therefore$ If his claim is justified. $X \sim \text{Bin}(6, \frac{5}{6})$

$$\begin{aligned} \therefore P(X \geq 4) &= P(X=4) + P(X=5) + P(X=6) \\ &= {}^6C_4 \left(\frac{5}{6}\right)^4 \left(\frac{1}{6}\right)^2 + {}^6C_5 \left(\frac{5}{6}\right)^5 \left(\frac{1}{6}\right)^1 + \left(\frac{5}{6}\right)^6 \\ &= 0.938 \text{ (8 d.p.)} \end{aligned}$$

(ii) If his claim is not justified and he throws an average just once in six throws then :-

$$X \sim \text{Bin}(6, \frac{1}{6})$$

$$\begin{aligned} P(X \geq 4) &= P(X=4) + P(X=5) + P(X=6) \\ &= {}^6C_4 \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^2 + {}^6C_5 \left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^1 + \left(\frac{1}{6}\right)^6 \\ &= 0.0087 \end{aligned}$$

H_0 : The man can throw a six, five times out of six throws

H_1 : The man can only throw a six, on average only once in six throws.

Ans/

$\therefore H_0$ is accepted if the mean can throw at least 4 sixes in 6 throws.

(a) $P(\text{accepting } H_0 | H_1 \text{ is true})$

$$= P(X \geq 4 | X \sim \text{Bin}(6, 1/6))$$

$$= 0.0087$$

$$\therefore P(\text{Type II error}) = 0.0087$$

(b) $P(\text{Rejecting } H_0 | H_0 \text{ is true}) = P(X < 4 | X \sim \text{Bin}(6, 5/6))$

$$= 1 - P(X \geq 4)$$

$$= 1 - 0.958$$

$$= 0.042$$

$$\therefore P(\text{Type I error}) = 0.062$$

196/

STATISTICAL TESTS FOR

WILCOXON SIGN TEST - page 2

SIGN TEST - page 3

CHI SQUARE - page 7

MANN WHITNEY UTEST - page 4

PEARSON'S PRODUCT MOMENT - page 11

SPEARMAN'S RHO - page 10

Operation schedule 7. The Wilcoxon matched-pairs signed ranks test

Data requirements

Scores must be paired off in some way, and of at least ordinal status.

Example

The Wilcoxon test can be carried out on the rating scores obtained for the two brands of washing up liquid, Gresego and Kwiklene (see pages 68–70). The null hypothesis (H_0) is that the two sets of scores do not differ, and the experimental hypothesis (H_1 or H_A) is that Gresego gives better results than Kwiklene.

Subject	Rating		Subject	Rating	
	Gresego	Kwiklene		Gresego	Kwiklene
1	8	5	6	7	6
2	7	5	7	9	5
3	9	2	8	6	5
4	7	6	9	5	6
5	8	9			

Step 1: The data will be paired off, but in doing so, it does not matter which column comes first. Subtract each number in the first column from its partner in the second. Record the differences, being sure to note the sign.

$$8 - 5 = +3, 7 - 5 = +2, \dots, 5 - 6 = -1$$

The differences are: +3, +2, +7, +1, -1, +1, +4, +1, -1

Step 2: Rank the differences according to size, and giving the smallest difference rank 1. Do not rank any values of 0 which occur. Ignore the signs during the ranking procedure. (See schedule 6 for ranking instructions.)

Differences +3 +2 +7 +1 -1 +1 +4 +1 -1
Ranks 7 6 9 3 3 3 8 3 3

Step 3: Add the ranks for all the differences which are positive, and all the negative differences. Keep the two totals separate.

Positive difference ranks 7 + 6 + 9 + 3 + 3 + 8 + 3 = 39

Negative difference ranks 3 + 3 = 6

Step 4: Whichever total of the two obtained in Step 3 is the smaller is the value of the test statistic T .

$$T = 6$$

Step 4: To determine N , in order to decide whether a given value of T is significant using table S2, count the number of pairs of scores used, but subtract any pairs whose difference was found to be 0 in Step 1.

There are 9 pairs of scores, and no values of 0. $N = 9$.

Operation schedule 7a. Evaluation of Wilcoxon's T statistic

Example

The data analysed using Wilcoxon's test, in schedule 7 will be evaluated for significance, using table S2. From the data, $T = 6$, and $N = 9$.

Step 1: Decide whether the experimental hypothesis (H_1) was a one-tailed or two-tailed hypothesis.

In the experiment on washing up liquid, it had been hypothesised that scores from Gresego

Table S2. Critical values of T for the Wilcoxon matched-pairs signed ranks test. T must be equal to or less than the stated value to be significant

N	Level of significance for one-tailed test				Level of significance for one-tailed test				
	0.05	0.025	0.01	0.005	0.05	0.025	0.01	0.005	
	Level of significance for two-tailed test				Level of significance for two-tailed test				
N	0.10	0.05	0.02	0.01	N	0.10	0.05	0.02	0.01
5	1	—	—	—	28	130	117	101	92
6	2	1	—	—	29	141	127	111	100
7	4	2	0	—	30	152	137	120	109
8	6	4	2	0	31	163	148	130	118
9	8	6	3	2	32	175	159	141	128
10	11	8	5	3	33	188	171	151	138
11	14	11	7	5	34	201	183	162	149
12	17	14	10	7	35	214	195	174	160
13	21	17	13	10	36	228	208	186	171
14	26	21	16	13	37	242	222	198	183
15	30	25	20	16	38	256	235	211	195
16	36	30	24	19	39	271	250	224	208
17	41	35	28	23	40	287	264	238	221
18	47	40	33	28	41	303	279	252	234
19	54	46	38	32	42	319	295	267	248
20	60	52	43	37	43	336	311	281	262
21	68	59	49	43	44	353	327	297	277
22	75	66	56	49	45	371	343	313	292
23	83	73	62	55	46	389	361	329	307
24	92	81	69	61	47	408	379	345	323
25	101	90	77	68	48	427	397	362	339
26	110	98	85	76	49	446	415	380	356
27	120	107	93	84	50	466	434	398	373

Source: F. Wilcoxon and R. A. Wilcox, *Some rapid approximate statistical procedures*, page 28, table 2, 1964. New York: Lederle Laboratories. Reproduced with the permission of American Cyanamid Company

Step 2: Locate the appropriate value for N in table S2, using the column headed N . Opposite $N = 9$, you should read the values 8 6 3 2.

Step 3: Move along the row until you arrive at the tabled value which is equal to or just larger than the value of T which you have just obtained.

The second value of 6 is equal to the T just obtained.

Step 4: Read the appropriate significance level off from the top of the column reached in Step 3. Take the one- or two-tailed probability level stated, according to the decision made in Step 1.

The obtained value of 6 is significant at the 0.025 level for a one-tailed hypothesis.

Step 5: State the conclusion.

The results were significant at the $p = 0.025$ level, using Wilcoxon's test ($N = 9$, $T = 6$; one-tailed hypothesis). Therefore the null hypothesis can be rejected, and it is concluded that the performance of Gresego was superior to that of Kwiklene.

Note

If the obtained value of T falls between tabled values, both of which are significant, then the most conservative (i.e. the least significant), and with the highest value of p)



Operation schedule 8: The sign test

Data requirements

The scores must be paired off in some way, but can be the simplest type of data (i.e. nominal).

Example

The data derived from the washing up experiment and analysed in schedule 7 will be used once more.

Subject	Rating		Subject	Rating	
	Gresgo	Kwiklenc		Gresgo	Kwiklenc
1	8	5	6	7	6
2	7	5	7	9	5
3	9	2	8	6	5
4	7	6	9	5	6
5	8	9			

Step 1: Put the data into a table, as above. It does not matter which column of scores comes first. Mentally subtract each value in the second column from its partner in the first. If the answer is positive, give the pairs a plus (+) sign, and if negative, a minus (-) sign. Note 0 if the two values are equal.

S	Gresgo	Kwiklenc	Sign	S	Gresgo	Kwiklenc	Sign
1	8	5	+	6	7	6	+
2	7	5	+	7	9	5	+
3	9	2	+	8	6	5	+
4	7	6	+	9	5	6	-
5	8	9	-				

8 The sign test

Step 2: Count the number of times the less frequent sign occurs, to obtain the statistic S .

The minus signs are less frequent, and occur twice. So $S = 2$.

Step 3: Count the total number of pluses and minuses, to obtain N . (Note that any differences of 0 will not be included in N .)

There are 9 signs, and so $N = 9$.

Step 4: Decide whether your test comprises a one- or two-tailed test.

The hypothesis under consideration is a one-tailed hypothesis.

Table S3. Critical values of S for the sign test. S must be equal to or less than the stated value to be significant

N	Level of significance for one-tailed test			Level of significance for two-tailed test		
	0.05	0.025	0.01	0.05	0.025	0.01
5	0	—	—	—	—	—
6	0	0	—	—	—	—
7	0	0	0	—	—	—
8	1	0	0	0	0	—
9	1	1	0	0	0	—
10	1	1	0	0	0	0
11	2	1	1	1	1	0
12	2	2	1	1	1	0
13	3	2	1	1	1	0
14	3	2	2	2	1	0
15	3	3	2	2	2	1
16	4	3	2	2	2	1
17	4	4	3	2	2	1
18	5	4	3	3	3	1
19	5	4	4	3	3	1
20	5	5	4	4	3	2
25	7	7	6	5	5	4
30	10	9	8	7	7	5
35	12	11	10	9	9	7

Step 5: Use table S3 to evaluate the statistic S (from Step 2), in conjunction with N (determined in Step 3). The table is used in just the same manner as table S2 for the Wilcoxon test (see schedule 7a).

Opposite the value 9, we read 1, 1, 0, 0, —. Our obtained value of $S = 2$ exceeds all these tabulated values, and so we conclude that as the probability exceeds that given for the 0.05 level, one-tailed test, our results are not significant.

Step 6: State the conclusion.

The results of the analysis were nonsignificant ($S = 2$, $N = 9$, sign test), and so the null hypothesis cannot be rejected. It is concluded that the two washing up liquids do not differ in their performance.

Operation schedule 9. The Mann-Whitney U test

Data requirements

Scores need to be of at least ordinal status.

Example

The results gathered whilst investigating a new memorising technique, described in chapter 1, will be used for analysis. They are:

Set 1: 30, 35, 45, 50, 75, 80 (the unaided group)

Set 2: 45, 50, 58, 62, 69, 70 (the aided group)

The experimental hypothesis was two-tailed, and the null hypothesis states that there will be no difference between the two sets of scores.

Step 1: Put the two sets of scores into two vertical columns, with a good gap between them. If one of the lists is smaller than the other, put it first, and call it 'list A'. If the lists are the same size, then it doesn't matter which comes first. Call the second list 'list B'.

List A	List B
30	45
35	50
45	58
50	62
75	69
80	70

Step 2: Count the number of scores in each list, to obtain N_A and N_B . Then multiply the two values together. This gives $N_A N_B$.

$$N_A = 6, N_B = 6; N_A N_B = 6 \times 6 = 36$$

Step 3: Take the number of scores in list A, and add the value of 1.

$$6 + 1 = 7$$

Step 4: Multiply the value found in Step 3 by the number of scores in list A, and divide the answer by 2.

$$\frac{7 \times 6}{2} = \frac{42}{2} = 21$$

Step 5: Rank all the numbers in both groups (see schedule 6 for ranking procedure), taking both sets of numbers together, and giving the smallest score rank 1. Write the ranks out in two more columns, headed R_A and R_B , and put each to the immediate right of the original lists of scores A and B.

List A	R_A	List B	R_B
30	1	45	31
35	2	50	51
45	31	58	7
50	51	62	8
75	11	69	9
80	12	70	10



Step 6: Add the ranks which were given to the items on list A (i.e. R_A)

$$1 + 2 + 31 + 51 + 11 + 12 = 35$$

Step 7: Add the value found in Step 2 to that found in Step 4, and subtract the value of Step 6.

$$36 + 21 - 35 = 22$$

Step 8: Subtract the value of Step 7 from the result of Step 2.

$$36 - 22 = 14$$

Step 9: The values of Steps 7 and 8 give two values for the statistics U and U' (pronounced 'U prime'). The smaller value will be U , and the larger U' .

$$U = 14 \text{ and } U' = 22$$

Step 10: Use the two values of N_A and N_B , found in Step 2, in conjunction with table S4 (overleaf), to determine whether the obtained value of U is significant. If U is equal to or smaller than the critical values listed, then the null hypothesis can be rejected. Two significance levels are given in the table, for the 0.05 and 0.01 probabilities (two-tailed), and the 0.025 and 0.005 probabilities (one-tailed tests).

Looking in the table at the point where $N_A = 6$ and $N_B = 6$ intersect, we read off the two critical values of 5 and 2. Our value of U (14) exceeds both numbers, and so we cannot reject the null hypothesis.

Step 11: State the conclusion.

The results of the analysis were nonsignificant ($U = 14$, $N_A = N_B = 6$; Mann-Whitney U test), and so the null hypothesis cannot be rejected. It is concluded that the scores obtained from subjects who were using the memory aid do not differ from those obtained by subjects without the aid.

Note

A useful check is to rework all the calculations, but call the original list A, list B. You should obtain precisely the same values for U and U' . Again, the smaller of the two values is labelled U , and evaluated.

Formulae

The formula for U is either

$$N_A N_B + \frac{N_A(N_A + 1)}{2} - \Sigma R_A$$

$$\text{or } N_A N_B + \frac{N_B(N_B + 1)}{2} - \Sigma R_B$$

where N_A = the number of items in list A

N_B = the number of items in list B

R_A = the sum of the ranks given to items in list A

R_B = the sum of the ranks given to items in list B

Table 5d Critical values of U for a one-tailed test at 0.05; two-tailed test at 0.10* (Mann-Whitney)

n_2	n_1																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
2	—	—	—	—	0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
3	—	—	0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
4	—	—	0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5	—	—	0	1	2	4	5	6	8	11	12	13	15	16	18	19	20	22	23	25
6	—	—	0	2	3	5	7	10	12	14	16	17	19	21	23	25	26	28	30	32
7	—	—	0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37
8	—	—	0	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44
9	—	—	1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51
10	—	—	1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58
11	—	—	1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65
12	—	—	2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72
13	—	—	2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80
14	—	—	2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87
15	—	—	3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	92
16	—	—	3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	100
17	—	—	3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	107
18	—	—	4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	115
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138

* Dashes in the body of the table indicate that no decision is possible at the stated level of significance.

For any n_1 and n_2 the observed value of U is significant at a given level of significance if it is equal to or less than the critical values shown.

SOURCE: R. Runyon and A. Haber (1976) Fundamentals of Behavioural Statistics (3rd ed.) Reading, Mass.: McGraw Hill, Inc. with kind permission of the publisher.

Table 5c Critical values of U for a one-tailed test at 0.025; two-tailed test at 0.05* (Mann-Whitney)

n_2	n_1																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	0	0	0	0	1	1	1	1	1	2	2	2	2
3	—	—	—	—	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	—	—	—	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	—	—	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	—	—	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	—	—	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	—	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	—	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	—	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	—	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	—	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	—	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	—	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	—	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	—	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	—	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	—	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	—	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	—	2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

* Dashes in the body of the table indicate that no decision is possible at the stated level of significance.

For any n_1 and n_2 the observed value of U is significant at a given level of significance if it is equal to or less than the critical values shown.

SOURCE: R. Runyon and A. Haber (1976) Fundamentals of Behavioural Statistics (3rd ed.) Reading, Mass.: McGraw Hill, Inc. with kind permission of the publisher.

Operation schedule 12. Simple chi-square

Data requirements

- 1 Data must be in frequency form, i.e. counted, rather than scores.
- 2 Entries in each cell must be independent.
- 3 The *expected* number for each cell must not be less than 5.

Example

One hundred and fifty-six boys and two hundred and four girls were asked whether they liked vanilla slices. Ninety-four boys said that they did, and one hundred and seventy-five girls. The remainder did not like them. Is there any evidence that there is a sex difference in preference for vanilla slices?

Step 1: Plot the data into a large 2×2 table. The number in each cell is the *observed* frequency. Obtain row and column totals, and a grand total (N).

	Column 1 Like slices	Column 2 Dislike slices	Totals
Row 1 Boys	Cell A 94	Cell B 62	156
Row 2 Girls	Cell C 175	Cell D 29	204
Totals	269	91	360

Step 2: Multiply the row 1 total with the column 1 total, and divide the answer by the overall total, N . This gives the *expected* frequency for cell A.

By multiplying the row and column totals relating to each cell together, and dividing the answer by the overall total, N , you obtain the *expected* frequencies for each cell. Write the value obtained for each cell below the *observed* frequency (with a different coloured pen).

$$\text{Cell A: } \frac{156 \times 269}{360} = 116.6$$

$$\text{Cell B: } \frac{156 \times 91}{360} = 39.4$$

$$\text{Cell C: } \frac{204 \times 269}{360} = 152.4$$

$$\text{Cell D: } \frac{204 \times 91}{360} = 51.6$$

Step 3: Find the difference between each *observed* and *expected* frequency for each cell, always taking the smaller from the larger. Then, for each cell value, subtract 0.5.

$$\text{Cell A: } 116.6 - 94 - 0.5 = 22.1$$

$$\text{Cell B: } 62 - 39.4 - 0.5 = 22.1$$

$$\text{Cell C: } 175 - 152.4 - 0.5 = 22.1$$

$$\text{Cell D: } 51.6 - 29 - 0.5 = 22.1$$

Step 4: Square each of the cell values obtained in Step 3, and divide the answer by the *expected* frequency for that particular cell.

$$\text{Cell A: } \frac{(22.1)^2}{116.6} = 4.189 \quad \text{Cell B: } \frac{(22.1)^2}{39.4} = 12.396$$

$$\text{Cell C: } \frac{(22.1)^2}{152.4} = 3.205 \quad \text{Cell D: } \frac{(22.1)^2}{51.6} = 9.465$$

Step 5: Add together the four values obtained in Step 4 to get χ^2 .

$$4.189 + 12.396 + 3.205 + 9.465 = 29.255$$

Step 6: Evaluate the value of χ^2 , using table S6. In a simple *chi*-square test the df will always be 1.

The obtained value of 29.25 exceeds the tabulated value of 10.83 which is significant at the 0.001 level (two-tailed test). The null hypothesis can be rejected.

Step 7: State the conclusion.

A *chi*-square test carried out on the data obtained was significant at the 0.001 level ($\chi^2 = 29.254$, $df = 1$), and so it is concluded that there is a difference between the sexes in liking for vanilla slices.

Formula

The formula for *chi*-square, in which Yates' correction is incorporated, is

$$\chi^2 = \sum \frac{(O - E - 0.5)^2}{E}$$

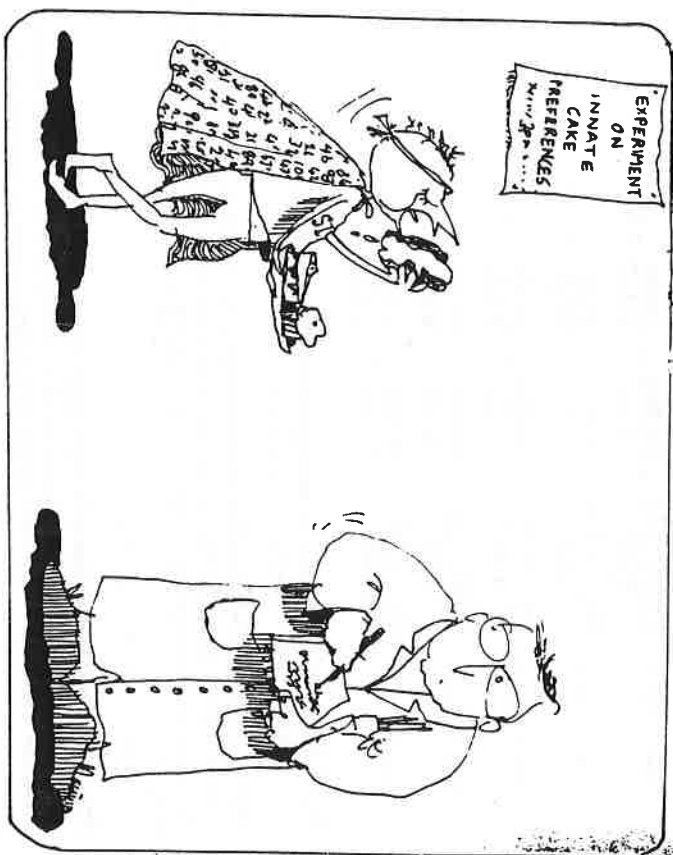
where E = the *expected* frequencies

O = the *observed* frequencies

Σ = the sum of

Table S6. Critical values of χ^2 . χ^2 must be equal to or more than the stated value to be significant

df	Level of significance for one-tailed test			Level of significance for two-tailed test		
	0.05	0.025	0.005	0.01	0.001	
1	2.706	3.841	6.635	10.83		
2	4.605	5.991	9.210	13.82		
3	6.251	7.815	11.34	16.27		
4	7.779	9.488	13.28	18.47		
5	9.236	11.07	15.09	20.52		
6	10.64	12.59	16.81	22.46		
7	12.02	14.07	18.48	24.32		
8	13.36	15.51	20.09	26.12		
9	14.68	16.92	21.67	27.88		
10	15.99	18.31	23.21	29.59		
11	17.28	19.68	24.73	31.26		
12	18.55	21.03	26.22	32.91		
13	19.81	22.36	27.69	34.53		



NOT ALL EXPERIMENTS ARE UNPLEASANT FOR THE SUBJECT...

Operation schedule 13. Complex chi-square

Data requirements

Data must be in frequency form, i.e. counted, rather than actual scores. Entries in each cell must be independent.

The expected number for each cell must not be less than 5.

Example

The different degree classifications obtained by finalists at Wetwang University (page 99) provide material for a complex chi-square analysis. The null hypothesis states that there is no difference between the three groups in terms of the proportions obtaining the different classifications.

Step 1: Put the data for the three groups of students and five degree classes into a 3×5 contingency table. The numbers in each cell stand for the numbers of students obtaining the particular degree and degree class, and are the *observed frequencies*.

Add together the numbers *across* the rows to get row totals, and *down* the columns to give column totals. The row and column totals added together should agree, and that total will be the overall total, N . Leave plenty of space in the table, for you will later be writing the expected values in each cell.

Subjects	Degree classification					Totals
	I	II	III	III	Pass	
Languages	5	10	20	35	30	100
Maths	35	40	80	25	20	200
Economics	0	10	10	20	20	60
Totals	40	60	110	80	70	360

Step 2: To compute the *expected frequency* for individual cells, follow this procedure. Take the row sum and the column sum which intersect at the cell, and multiply the two values together. Then divide the answer by the grand total, N .

For the top left-hand cell, Languages, and a class I degree, it will be

$$\text{Row 1, Col. 1: } \frac{100 \times 40}{360} = \frac{4000}{360} = 11.111$$

Don't round off the decimal places at less than three.

$$\text{Row 2, Col. 1: } \frac{200 \times 40}{360} = 22.222$$

$$\text{Row 3, Col. 1: } \frac{60 \times 40}{360} = 6.667$$

$$\text{Row 1, Col. 2: } \frac{100 \times 60}{360} = 16.667$$

$$\text{Row 2, Col. 2: } \frac{200 \times 60}{360} = 33.333$$

$$\text{Row 3, Col. 2: } \frac{60 \times 60}{360} = 10.000$$

$$\text{Row 1, Col. 3: } \frac{100 \times 110}{360} = 30.556$$

$$\text{Row 2, Col. 3: } \frac{200 \times 110}{360} = 61.111$$

$$\text{Row 3, Col. 3: } \frac{60 \times 110}{360} = 18.333$$

$$\text{Row 1, Col. 4: } \frac{100 \times 80}{360} = 22.222$$

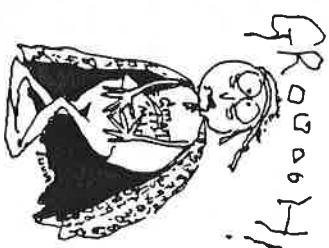
$$\text{Row 2, Col. 4: } \frac{200 \times 80}{360} = 44.444$$

$$\text{Row 3, Col. 4: } \frac{60 \times 80}{360} = 13.333$$

$$\text{Row 1, Col. 5: } \frac{100 \times 70}{360} = 19.444$$

$$\text{Row 2, Col. 5: } \frac{200 \times 70}{360} = 38.889$$

$$\text{Row 3, Col. 5: } \frac{60 \times 70}{360} = 11.667$$



... WELL, NOT AT THE TIME OF PRACTICATION

Step 3: When the expected values have been calculated, put each one into the table below the appropriate obtained frequency. Check that they all exceed 4.

Subjects	Degree classification				Totals
	I	II	III	Pass	
Languages	5	10	20	35	100
	11.111	16.667	30.556	22.222	19.444
Maths	35	40	80	25	200
	22.222	33.333	61.111	44.444	38.889
Economics	0	10	10	20	60
	6.667	10.000	18.333	13.333	11.667
Totals	40	60	110	80	360

Step 4: To obtain the values which will be added together to get χ^2 , the following procedure must be undertaken for each cell. The difference between the obtained and expected frequency for every cell is found, i.e. by subtracting the smaller value from the larger. This is then squared, and the value divided by the expected frequency for that particular cell.

Row 1, Col. 1:	$\frac{(11.111 - 5)^2}{11.111} = \frac{(6.111)^2}{11.111} = 3.344$	$\frac{(6.111)^2}{11.111} = 3.361$
Row 2, Col. 1:	$\frac{(35 - 22.222)^2}{22.222} = \frac{(12.778)^2}{22.222} = 163.277$	$\frac{(12.778)^2}{22.222} = 7.348$
Row 3, Col. 1:	$\frac{(6.667 - 0)^2}{6.667} = \frac{(6.667)^2}{6.667} = 6.667$	$\frac{(6.667)^2}{6.667} = 6.667$
Row 1, Col. 2:	$\frac{(16.667 - 10)^2}{16.667} = \frac{(6.667)^2}{16.667} = 44.449$	$\frac{(6.667)^2}{16.667} = 2.667$
Row 2, Col. 2:	$\frac{(40 - 33.333)^2}{33.333} = \frac{(6.667)^2}{33.333} = 44.449$	$\frac{(6.667)^2}{33.333} = 1.333$
Row 3, Col. 2:	$\frac{(10 - 10)^2}{10} = \frac{(0)^2}{10} = 0$	$\frac{(0)^2}{10} = 0$
Row 1, Col. 3:	$\frac{(30.556 - 20)^2}{30.556} = \frac{(10.556)^2}{30.556} = 111.429$	$\frac{(10.556)^2}{30.556} = 3.647$
Row 2, Col. 3:	$\frac{(80 - 61.111)^2}{61.111} = \frac{(18.889)^2}{61.111} = 356.794$	$\frac{(18.889)^2}{61.111} = 5.838$
Row 3, Col. 3:	$\frac{(18.333 - 10)^2}{18.333} = \frac{(8.333)^2}{18.333} = 69.439$	$\frac{(8.333)^2}{18.333} = 3.788$
Row 1, Col. 4:	$\frac{(35 - 22.222)^2}{22.222} = \frac{(12.778)^2}{22.222} = 163.277$	$\frac{(12.778)^2}{22.222} = 7.348$

13 Complex chi-square

Row 2, Col. 4:	$\frac{(44.444 - 25)^2}{44.444} = \frac{(19.444)^2}{44.444} = 8.507$	$\frac{(19.444)^2}{44.444} = 8.507$
Row 3, Col. 4:	$\frac{(20 - 13.333)^2}{13.333} = \frac{(6.667)^2}{13.333} = 3.334$	$\frac{(6.667)^2}{13.333} = 3.334$
Row 1, Col. 5:	$\frac{(30 - 19.444)^2}{19.444} = \frac{(10.556)^2}{19.444} = 5.731$	$\frac{(10.556)^2}{19.444} = 5.731$
Row 2, Col. 5:	$\frac{(38.889 - 20)^2}{38.889} = \frac{(18.889)^2}{38.889} = 9.175$	$\frac{(18.889)^2}{38.889} = 9.175$
Row 3, Col. 5:	$\frac{(20 - 11.667)^2}{11.667} = \frac{(8.333)^2}{11.667} = 5.952$	$\frac{(8.333)^2}{11.667} = 5.952$

Step 5: Obtain the value of χ^2 by adding all the values obtained in Step 4.
 $3.361 + 7.348 + 6.667 + \dots + 9.175 + 5.952 = 74.696$

Step 6: Obtain the df which will be needed to evaluate χ^2 . It will always be the number of rows minus 1 multiplied by the number of columns minus 1, unless you are using the steps for a one-sample χ^2 -square, when the df is always the number of cells minus 1.
 $df = (3 - 1)(5 - 1) = 2 \times 4 = 8$

Step 7: Use table S6 to evaluate the significance of the χ^2 value obtained.

When $df = 8$, we read off the values 13.36, 15.51, 20.09 and 26.12. Our value exceeds all these, and so we conclude that it is significant at the 0.001 level for a two-tailed test.

Step 8: State the conclusion.

The different degree classifications for finalists in languages, maths and economics were analysed, using χ^2 -square. The value of χ^2 was found to be significant at the 0.001 level ($\chi^2 = 74.696$, $df = 8$), and so it was concluded that the proportions of students obtaining the five classes of degree vary substantially between subjects.

Note

The test only tells us that the numbers come from distributions having different shapes. We have to decide how the shapes differ, and where the most substantial discrepancies lie by inspection of the data. In addition, take care over the interpretation of this significant result. We cannot say that 'brighter' students take maths etc., for the results may be dependent upon factors other than students' abilities, such as departmental marking policies. χ^2 -square is only a measure of association.

Formula

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where E = the expected frequencies

O = the observed frequencies

Σ = the sum of



*Maud has a garden of roses
And flies fast on a lawn;*

Operation schedule 14. Spearman's rho

Data requirements

- 1 Scores must be of at least ordinal status.
- 2 Scores for comparison must be paired off in some manner. Usually they will have been obtained from one source, but this may not always be the case.
- 3 When plotted on a scattergram, the points can be in a linear or slightly curved pattern.

Example

It was noted that residents at the Heyshott Asylum frequently referred to Tennyson's poem *Maud*. Enquiries revealed that those patients most seriously afflicted with insanity (as assessed by the Heyshott Lunacy Scale, HLS) appeared to have read the poem rather often. All inmates were given a rating on the HLS, and asked how many times they had read *Maud*. The results were then used to obtain Spearman's rho.

Subject	HLS score	Maud readings
1	10	0
2	100	4
3	125	4
4	45	15
5	90	23
6	170	45
7	165	61

Because the staff anticipated that degree of insanity would be positively associated with number of *Maud* readings, they have established a one-tailed prediction.

Step 1: Draw a scattergram for the two sets of data; the vertical axis measuring one variable, and the horizontal axis the other.

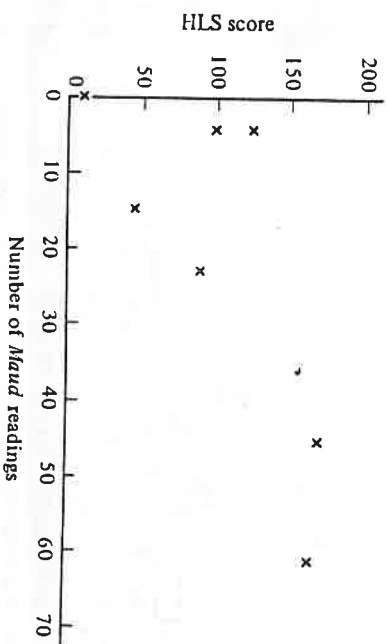


Figure 1. Scattergram showing the relationship between HLS scores and number of *Maud* readings

Step 2: Now cast the data into a table with seven columns. These will be:

- A The label for each subject, animal or source of paired scores, using numbers or letters of the alphabet.
- B The scores on one variable.
- C The scores on the second variable.
- D The scores from A, given ranks (see schedule 6).
- E The scores from B, given ranks.
- F Each value in list R_B subtracted from its partner in R_A .
- G Each value in D squared.

S	A	B	R_A	R_B	D	D^2
1	10	0	1	1	0	0
2	100	4	4	2	1	2
3	125	4	5	2	2	4
4	45	15	2	4	-2	4
5	90	23	3	5	-2	4
6	170	45	7	6	1	1
7	165	61	6	7	-1	1

Step 3: Count the number of paired scores in the sample, to obtain N.

There are 7 pairs of scores; $N = 7$.

Step 4: Multiply N by its own value, twice, and then subtract its own value.

$$(7 \times 7 \times 7) - 7 = 343 - 7 = 336$$

Step 5: Total the values in column D^2 .

$$0 + 2 + 4 + 4 + 4 + 1 + 1 = 18$$

Step 6: Multiply the value obtained in Step 5 by the number 6; then divide the result by the value found in Step 4.

$$\frac{18 \times 6}{336} = \frac{111}{336} = 0.3304$$

Step 7: To find rho, subtract the value of Step 6 from the number 1. Retain the sign. The result should always lie between -1 and +1.

$$1 - 0.3304 = +0.6696$$

Step 8: Evaluate the significance of rho, using table S7. Follow the numbers along the row opposite the appropriate N, to find the value of rho which must be exceeded for the various significance levels given.

When $N = 7$, we read 0.714, 0.786, 0.893 and 0.929 for the one-tailed significance levels. Our obtained value of rho, 0.6696, does not exceed any of these, and so the null hypothesis cannot be rejected.

Step 9: State the conclusion.

From the results of a correlation carried out on the number of times a patient had read the poem *Maud*, and their rating on the HLS, it was found that there did not appear to be a significant association between the two variables ($\rho = +0.67, N = 7$). However, the association was positive, and the relationship may be found to be a significant one if a larger sample were used.

Formula and abbreviations

Spearman's ρ is sometimes written ρ or r_s . The formula used for calculations is

$$r_s = 1 - \frac{6\sum D^2}{(N^3 - N)}$$

where $\sum D^2$ = the squared values of the differences between the ranked scores, totalled
 N = the number of paired scores

Table S7: Critical values of Spearman's rho. Rho must be equal to or more than the stated value to be significant

N	Level of significance for one-tailed test				Level of significance for a two-tailed test			
	0.05	0.025	0.01	0.005	0.1	0.05	0.02	0.01
5	0.900	1.000	1.000	—	0.564	0.648	0.746	0.794
6	0.829	0.886	0.943	1.000	0.506	0.591	0.712	0.777
7	0.714	0.786	0.893	0.929	0.456	0.544	0.645	0.715
8	0.643	0.738	0.833	0.881	0.425	0.506	0.601	0.665
9	0.600	0.683	0.783	0.833	0.399	0.475	0.564	0.625
10	0.564	0.648	0.746	0.794	0.377	0.450	0.534	0.591
12	0.506	0.591	0.712	0.777	0.359	0.428	0.508	0.562
14	0.456	0.544	0.645	0.715	0.343	0.409	0.485	0.537
16	0.425	0.506	0.601	0.665	0.329	0.392	0.465	0.515
18	0.399	0.475	0.564	0.625	0.317	0.377	0.448	0.496
20	0.377	0.450	0.534	0.591	0.306	0.364	0.432	0.478
22	0.359	0.428	0.508	0.562				
24	0.343	0.409	0.485	0.537				
26	0.329	0.392	0.465	0.515				
28	0.317	0.377	0.448	0.496				
30	0.306	0.364	0.432	0.478				

N = the number of paired scores used.
Treat a negative value of rho as if it were positive, when using the table, but when interpreting it don't forget that it will indicate an inverse relationship. Source: Snodgrass, table C.6

Operation schedule 15. The Pearson product-moment correlation (r)

Data requirements

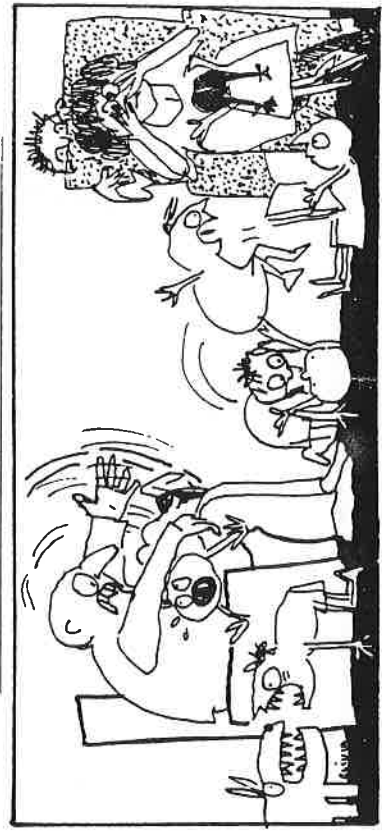
- 1 Scores for comparison must be paired off in some manner. Usually they will have been obtained from the same source, but this may not always be the case.
- 2 The relationship between the two variables must be a linear one. This can be assessed from a scattergram.
- 3 The scores must be of at least interval status.
- 4 The scores must be normally distributed.
- 5 The two sets of scores must have similar variances.

Example

A schoolgirl, recently annoyed by the antics of her small sisters, vowed that she would never inflict siblings on any child she might have when she was older. Curious to know her classmates' opinions, she asked six of them how many children they thought was a good number to have, and compared this with the number of children under the age of fifteen actually in each one's family.

Bearing in mind her own experiences, she imagined that the more siblings a person had, the fewer they would think was desirable! Thus she has predicted a negative correlation: this directional prediction constitutes a one-tailed hypothesis. The data she collected are given below:

Subject	No. of children in own family	No. of children thought ideal
Christine	1	5
David	2	3
Eric	3	4
Lynne	4	2
Margaret	4	1
Tom	5	1



MOST PARENTS HOLD THE SAME VIEW ABOUT THE IDEAL NUMBER OF CHILDREN

Operation schedules

Step 1: Draw a scattergram for the two sets of data, the vertical axis measuring one variable, the horizontal the other. At this stage, check for linearity.

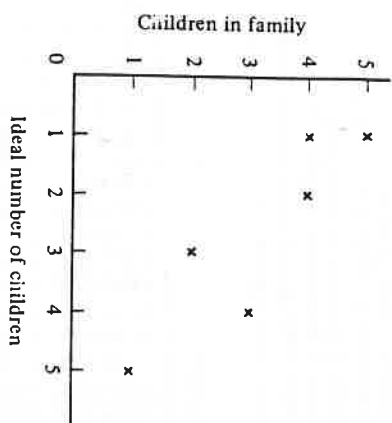


Figure 2. Scattergram showing the relationship between the number of children in a young person's family and the number that they would consider to be 'ideal'.

Step 2: Now cast the data into a table, with six columns. These will be:

5 The label for each subject, animal or source of paired scores, using numbers or letters of the alphabet.

A The scores on one variable (and which are added to give ΣA).

A² Each A score squared (and which added gives ΣA^2).

B The scores on the second variable (and which are added to give ΣB).

B² Each B score squared (and which added gives ΣB^2).

AB Each A score multiplied by its matching B score. (Added gives ΣAB)

Add all the columns, apart from the first one. This is a good stage to check that the data meets all the requirements other than linearity.

Subject	A	A ²	B	B ²	AB
C	1	1	5	25	5
D	2	4	3	9	6
E	3	9	4	16	12
F	4	16	2	4	8
G	5	25	1	1	5
H	1	1	1	1	1
I	2	4	2	4	4
J	3	9	3	9	9
K	4	16	4	16	16
L	5	25	5	25	25
M	1	1	1	1	1
N	2	4	2	4	4
O	3	9	3	9	9
P	4	16	4	16	16
Q	5	25	5	25	25
R	1	1	1	1	1
S	2	4	2	4	4
T	3	9	3	9	9
U	4	16	4	16	16
V	5	25	5	25	25
W	1	1	1	1	1
X	2	4	2	4	4
Y	3	9	3	9	9
Z	4	16	4	16	16
AA	5	25	5	25	25
AB	1	1	1	1	1
AC	2	4	2	4	4
AD	3	9	3	9	9
AE	4	16	4	16	16
AF	5	25	5	25	25
AG	1	1	1	1	1
AH	2	4	2	4	4
AI	3	9	3	9	9
AJ	4	16	4	16	16
AK	5	25	5	25	25
AL	1	1	1	1	1
AM	2	4	2	4	4
AN	3	9	3	9	9
AO	4	16	4	16	16
AP	5	25	5	25	25
AQ	1	1	1	1	1
AR	2	4	2	4	4
AS	3	9	3	9	9
AT	4	16	4	16	16
AU	5	25	5	25	25
AV	1	1	1	1	1
AW	2	4	2	4	4
AX	3	9	3	9	9
AY	4	16	4	16	16
AZ	5	25	5	25	25
BA	1	1	1	1	1
BB	2	4	2	4	4
BC	3	9	3	9	9
BD	4	16	4	16	16
BE	5	25	5	25	25
BF	1	1	1	1	1
BG	2	4	2	4	4
BH	3	9	3	9	9
BI	4	16	4	16	16
BJ	5	25	5	25	25
BK	1	1	1	1	1
BL	2	4	2	4	4
BM	3	9	3	9	9
BN	4	16	4	16	16
BO	5	25	5	25	25
BP	1	1	1	1	1
BQ	2	4	2	4	4
BR	3	9	3	9	9
BS	4	16	4	16	16
BT	5	25	5	25	25
BU	1	1	1	1	1
BV	2	4	2	4	4
BW	3	9	3	9	9
BX	4	16	4	16	16
BY	5	25	5	25	25
BZ	1	1	1	1	1
CA	2	4	2	4	4
CB	3	9	3	9	9
CC	4	16	4	16	16
CD	5	25	5	25	25
CE	1	1	1	1	1
CF	2	4	2	4	4
CG	3	9	3	9	9
CH	4	16	4	16	16
CI	5	25	5	25	25
CJ	1	1	1	1	1
CK	2	4	2	4	4
CL	3	9	3	9	9
CM	4	16	4	16	16
CN	5	25	5	25	25
CO	1	1	1	1	1
CP	2	4	2	4	4
CQ	3	9	3	9	9
CR	4	16	4	16	16
CS	5	25	5	25	25
CT	1	1	1	1	1
CU	2	4	2	4	4
CV	3	9	3	9	9
CW	4	16	4	16	16
CX	5	25	5	25	25
CY	1	1	1	1	1
CZ	2	4	2	4	4
DA	3	9	3	9	9
DB	4	16	4	16	16
DC	5	25	5	25	25
DD	1	1	1	1	1
DE	2	4	2	4	4
DF	3	9	3	9	9
DG	4	16	4	16	16
DH	5	25	5	25	25
DI	1	1	1	1	1
DJ	2	4	2	4	4
DK	3	9	3	9	9
DL	4	16	4	16	16
DM	5	25	5	25	25
DN	1	1	1	1	1
DO	2	4	2	4	4
DP	3	9	3	9	9
DQ	4	16	4	16	16
DR	5	25	5	25	25
DS	1	1	1	1	1
DT	2	4	2	4	4
DU	3	9	3	9	9
DV	4	16	4	16	16
DW	5	25	5	25	25
DX	1	1	1	1	1
DY	2	4	2	4	4
DZ	3	9	3	9	9
EA	4	16	4	16	16
EB	5	25	5	25	25
EC	1	1	1	1	1
ED	2	4	2	4	4
EE	3	9	3	9	9
EF	4	16	4	16	16
EG	5	25	5	25	25
EH	1	1	1	1	1
EI	2	4	2	4	4
EJ	3	9	3	9	9
EK	4	16	4	16	16
EL	5	25	5	25	25
EM	1	1	1	1	1
EN	2	4	2	4	4
EO	3	9	3	9	9
EP	4	16	4	16	16
EQ	5	25	5	25	25
ER	1	1	1	1	1
ES	2	4	2	4	4
ET	3	9	3	9	9
EU	4	16	4	16	16
EV	5	25	5	25	25
EW	1	1	1	1	1
EX	2	4	2	4	4
EY	3	9	3	9	9
EZ	4	16	4	16	16
FA	5	25	5	25	25
FB	1	1	1	1	1
FC	2	4	2	4	4
FD	3	9	3	9	9
FE	4	16	4	16	16
FF	5	25	5	25	25
FG	1	1	1	1	1
FH	2	4	2	4	4
FI	3	9	3	9	9
FJ	4	16	4	16	16
FK	5	25	5	25	25
FL	1	1	1	1	1
FM	2	4	2	4	4
FN	3	9	3	9	9
FO	4	16	4	16	16
FP	5	25	5	25	25
FO	1	1	1	1	1
FP	2	4	2	4	4
FO	3	9	3	9	9
FP	4	16	4	16	16
FO	5	25	5	25	25
FP	1	1	1	1	1
FO	2	4	2	4	4
FP	3	9	3	9	9
FO	4	16	4	16	16
FP	5	25	5	25	25
FO	1	1	1	1	1
FP	2	4	2	4	4
FO	3	9	3	9	9
FP	4	16	4	16	16
FO	5	25	5	25	25
FP	1	1	1	1	1
FO	2	4	2	4	4
FP	3	9	3	9	9
FO	4	16	4	16	16
FP	5	25	5	25	25
FO	1	1	1	1	1
FP	2	4	2	4	4
FO	3	9	3	9	9
FP	4	16	4	16	16
FO	5	25	5	25	25
FP	1	1	1	1	1
FO	2	4	2	4	4
FP	3	9	3	9	9
FO	4	16	4	16	16
FP	5	25	5	25	25
FO	1	1	1	1	1
FP	2	4	2	4	4
FO	3	9	3	9	9
FP	4	16	4	16	16
FO	5	25	5	25	25
FP	1	1	1	1	1
FO	2	4	2	4	4
FP	3	9	3	9	9
FO	4	16	4	16	16
FP	5	25	5	25	25
FO	1	1	1	1	1
FP	2	4	2	4	4
FO	3	9	3	9	9
FP	4	16	4	16	16
FO	5	25	5	25	25
FP	1	1	1	1	1
FO	2	4	2	4	4
FP	3	9	3	9	9
FO	4	16	4	16	16
FP	5	25	5	25	25
FO	1	1	1	1	1
FP	2	4	2	4	4
FO	3	9	3	9	9
FP	4	16	4	16	16
FO	5	25	5	25	25
FP	1	1	1	1	1
FO	2	4	2	4	4
FP	3	9	3	9	9
FO	4	16	4	16	16
FP	5	25	5	25	25
FO	1	1	1	1	1
FP	2	4	2	4	4
FO	3	9	3	9	9
FP	4	16	4	16	16
FO	5	25	5	25	25
FP	1	1	1	1	1
FO	2	4	2	4	4
FP	3	9	3	9	9
FO	4	16	4	16	16
FP	5	25	5	25	25
FO	1	1	1	1	1
FP	2	4	2	4	4
FO	3	9	3	9	9
FP	4	16	4	16	16
FO	5	25	5	25	25
FP	1	1	1	1	1
FO	2	4	2	4	4
FP	3	9	3	9	9
FO	4	16	4	16	16
FP	5	25	5	25	25
FO	1	1	1	1	1
FP	2	4	2	4	4
FO	3	9	3	9	9
FP	4	16	4	16	16
FO	5	25	5	25	25
FP	1	1	1	1	1
FO	2	4	2	4	4
FP	3	9	3	9	9
FO	4	16	4	16	16
FP	5	25	5	25	25
FO	1	1	1	1	1
FP	2	4	2	4	4
FO	3	9	3	9	9
FP	4	16	4	16	16
FO	5	25	5	25	25
FP	1	1	1	1	1

Table S8 Critical values of Pearson's r. r must be equal to or more than the stated value to be significant

N - 2	Level of significance for one-tailed test			
	0.05	0.025	0.005	0.0005
	Level of significance for two-tailed test			
	0.10	0.05	0.01	0.001
2	0.9000	0.9500	0.9900	0.9999
3	0.805	0.878	0.9587	0.9911
4	0.729	0.811	0.9172	0.9741
5	0.669	0.754	0.875	0.9509
6	0.621	0.707	0.834	0.9241
7	0.582	0.666	0.798	0.898
8	0.549	0.632	0.765	0.872
9	0.521	0.602	0.735	0.847
10	0.497	0.576	0.708	0.823
11	0.476	0.553	0.684	0.801
12	0.457	0.532	0.661	0.780
13	0.441	0.514	0.641	0.760
14	0.426	0.497	0.623	0.742
15	0.412	0.482	0.606	0.725
16	0.400	0.468	0.590	0.708
17	0.389	0.456	0.575	0.693
18	0.378	0.444	0.561	0.679
19	0.369	0.433	0.549	0.665
20	0.360	0.423	0.537	0.652
25	0.323	0.381	0.487	0.597
30	0.296	0.349	0.449	0.554
35	0.275	0.325	0.418	0.519
40	0.257	0.304	0.393	0.490
45	0.243	0.288	0.372	0.465
50	0.231	0.273	0.354	0.443
60	0.211	0.250	0.325	0.408
70	0.195	0.232	0.302	0.380
80	0.183	0.217	0.283	0.357
90	0.173	0.205	0.267	0.338
100	0.164	0.195	0.254	0.321

Source: Powell, page 69

15 The Pearson product-moment correlation

Formula

The formula for Pearson's r is:

$$r = \frac{N\sum AB - (\sum A)(\sum B)}{\sqrt{[N\sum A^2 - (\sum A)^2][N\sum B^2 - (\sum B)^2]}}$$

where N = the number of pairs of scores

$\sum AB$ = the sum of the products of the paired scores.

$\sum A$ = the sum of the scores on one variable

$\sum B$ = the sum of the scores on the second variable

$\sum A^2$ = the sum of the squared scores on the A variable

$\sum B^2$ = the sum of the squared scores on the B variable

Table S9 Random numbers

19	90	70	99	00	20	21	14	68	86	14	52	41	52	48	87	63	91	95	17	11	29	01	95	80
65	97	38	20	46	85	43	01	72	73	03	37	18	39	11	08	61	74	31	69	89	74	39	82	15
51	67	11	52	49	59	97	50	99	52	18	16	36	78	86	08	52	85	08	40	87	80	61	65	31
17	95	70	45	80	72	68	49	29	31	56	80	30	19	44	89	85	84	46	06	59	73	19	85	23
63	52	01	41	88	02	84	27	83	78	35	34	08	74	42	29	72	23	19	66	56	45	65	79	
60	61	97	22	61	49	64	92	85	44	01	64	18	30	06	16	40	12	89	88	50	14	49	81	06
98	99	46	50	47	12	83	11	41	16	63	14	52	32	53	25	48	19	68	70	77	02	54	00	52
76	38	03	29	63	79	43	61	40	15	86	63	59	80	02	14	33	40	39	27	31	58	50	28	
53	05	70	53	30	38	06	38	21	01	47	59	38	00	14	27	47	07	20	54	96	87	53	32	
02	87	40	41	45	47	24	49	57	74	22	13	88	83	31	32	25	62	17	10	97	11	69	84	
35	14	97	35	33	68	95	23	92	33	56	54	29	56	93	87	02	22	57	51	61	09	43	95	06
94	51	32	41	97	13	79	93	77	55	14	44	99	81	07	39	77	32	77	09	85	52	05	30	62
94	51	32	41	97	13	79	93	77	55	14	44	99	81	07	39	77	32	77	09	85	52	05	30	62
65	09	29	15	03	20	44	90	32	64	53	89	74	60	41	97	67	63	99	61	46	38	03	93	22
20	71	53	20	53	73	37	32	04	05	56	07	93	89	30	69	30	16	09	05	88	69	58	28	99
01	82	77	12	07	10	63	76	35	19	48	56	27	44	87	03	04	79	88	08	13	13	85	51	
53	43	37	15	36	92	38	70	96	92	82	11	08	95	97	52	06	79	45	82	63	18	27	44	
40	36	40	96	76	93	86	52	77	65	99	82	93	24	98	15	33	59	05	28	22	87	26	07	47
99	63	22	32	98	18	46	23	34	27	43	11	79	99	31	85	11	99	24	44	49	18	09	79	49
58	24	82	03	47	24	53	63	94	09	74	54	13	26	94	21	10	76	47	91	44	04	95	49	66
47	83	51	62	74	22	06	34	72	63	02	32	92	08	99	82	21	15	65	20	33	29	94	71	11
23	05	47	45	07	16	39	33	66	18	55	63	77	09	98	50	10	56	79	77	21	30	27	12	
69	81	21	99	21	29	70	83	61	51	70	47	14	32	99	74	20	52	30	87	09	41	15	09	
35	07	44	75	47	57	90	12	02	07	54	96	09	11	00	23	47	37	17	31	54	08	01	88	63
55	34	57	72	60	33	35	72	67	47	82	80	84	25	39	77	34	55	45	70	08	18	27	38	90
69	66	92	10	69	49	41	31	06	70	05	98	90	07	35	42	38	06	45	18	64	73	31	65	
90	92	10	80	65	19	69	02	83	67	72	16	42	79	60	75	86	90	68	24	64	19	35	51	
86	96	98	29	98	09	49	38	76	43	63	49	30	21	00	22	00	27	69	85	29	81	94	78	70
74	18	32	23	02	98	77	87	68	07	66	39	67	98	60	91	51	67	62	44	40	98	05	93	78
39	60	04	59	81	00	41	86	79	79	47	53	38	00	68	47	22	20	35	55	31	51	51		
15	91	29	12	03	57	99	90	37	75	91	12	81	19	36	61	32	08	58	37	40	13	68	07	
90	49	22	03	61	12	59	52	57	02	55	65	79	78	07	22	07	90	47	03	28	14	11	30	79
98	60	16	03	03	31	51	10	96	46	54	34	81	85	35	92	06	88	07	77	56	11	50	81	69
39	41	88	92	10	96	11	83	44	80	03	92	18	66	75	34	68	35	48	77	33	42	40	90	60
16	95	86	70	75	85	47	04	66	08	00	83	26	01	03	34	72	57	59	13	82	43	80	46	15
52	53	37	07	15	72	83	32	99	90	06	66	24	12	28	63	95	73	76	63	89	73	44	99	05
56	61	87	39	12	91	36	74	43	53	13	39	54	10	28	30	82	13	54	00	78	45	98	35	
21	94	47	90	12	77	53	84	46	47	82	13	40	21	11	91	18	95	58	24	16	74	11	53	
23	34	65	41	18	37	37	39	19	65	65	80	39	07	84	70	07	48	53	21	40	06	71		
00	83	63	22	55	34	18	04	52	35	99	01	30	98	64	56	27	09	34	86	61	85	53	83	45
87	64	07	83	11	20	99	45	18	45	76	08	64	27	48	13	93	55	34	18	37	79	49	90	
20	69	22	40	98	27	37	83	28	71	69	62	03	42	73	00	06	41	74	45	89	09	39	84	
40	23	72	51	39	10	65	81	92	59	73	42	37	11	61	58	76	17	14	97	04	76	62	16	17
73	96	53	97	86	59	71	74	17	32	64	63	91	08	25	35	10	24	19	23	71	82	13	74	
38	26	61	70	04	33	73	99	19	87	95	60	78	46	75	26	72	39	32	67	53	77	68	93	
48	67	26	43	18	74	17	43	96	99	17	43	48	76	43	00	65	08	50	45	60	33	01	07	
55	03	36	67	68	72	87	08	62	40	62	01	61	16	06	10	89	20	23	21	34	74	07		
44	10	13	85	57	73	96	07	04	52	19	59	50	88	92	09	65	90	77	47	25	76	16	19	31
95	06	79	88	54	79	96	23	53	10	48	03	45	15	22	65	39	07	16	29	45	33	02	43	70
68	15	54	35	02	42	35	48	06	32	95	33	95	22	00	18	72	00	18	23	85	61	68	90	
58	42	36	72	24	58	37	52	18	51	90	84	60	79	80	24	36	59	47	38	67	80	43	79	33
95	67	47	20	83	04	69	40	06	07	40	40	62	98	82	54	97	20	56	95	27	62	50	96	72
98	57	07	23	69	65	95	39	09	58	20	31	89	03	43	38	46	82	68	72	33	78	80	87	15
56	69	47	07	41	90	22	91	07	12	71	59	73	05	08	22	23	77	13	13	92	66	99		

Source: Powell, page 55